THIS WEEK

EDITORIALS

Top

- **Desirable partners**

  US refusal to collaborate with China on space science is short-sighted and misguided, from both a scientific and a pragmatic standpoint.
- **Climate change**

  Negotiations in Durban over greenhouse-gas emissions should not try to revive Kyoto.
- **Ex factor**

  Demise of snails in a New Zealand freezer is a sign of the times.

WORLD VIEW

Top

- **Time to stop celebrating the polluters**

  The United Nations must include sustainability in its quality-of-life index to encourage countries to develop responsibly, says Chuluun Togtokh.

RESEARCH HIGHLIGHTS

Top

- **Climate change: Mediterranean drying**
- **Organic chemistry: Carbon dioxide conversions**
- **Neurodevelopment: Cell source for brain disorder**
- **Palaeontology: Ancient creature's surprising sight**
- **Geology: Earthquake risk has not risen**
- **Neuroimaging: Getting past a brain block**
- **Metabolism: Genetic switch for big muscles**
- **Animal cognition: Jays plan meals in advance**
- COMMUNITY CHOICE

  Developmental biology: How the zebrafish brain mends itself

SEVEN DAYS

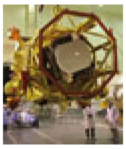
Top

- **Seven days: 11–17 November 2011**

The week in science: Geron stops clinical trials with human embryonic stem cells; NSF starts high-risk grants programme; and disgraced psychologist Stapel returns his PhD.

NEWS IN FOCUS

- **Russia gets the red planet blues**

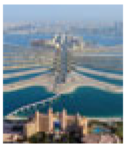
  Phobos probe failure puts planetary comeback in doubt.

  - Eric Hand
- **China forges ahead in space**

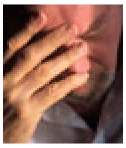
  Mars-probe problems are a minor blip in a bold strategy.

  - David Cyranoski
- **Gulf ecology hit by coastal development**

  Dubai's artificial islands are affecting marine ecosystems.

  - Daniel Cressey
- **Depression drug disappoints**

  Failure of a promising compound casts a shadow on others.

  - Heidi Ledford
- **Summit urged to clean up farming**

  Leading scientists say that agriculture is a 'poor relation' in global-warming negotiations.

- Natasha Gilbert

- **Targeted treatment tested as potential cancer cure**

Trial will deploy genetically targeted therapy early, rather than as last resort.

- Erika Check Hayden

- **Iran’s nuclear plan revealed**

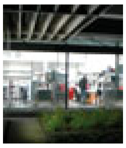
Report paints detailed picture of nation’s intention to build a warhead.

- Geoff Brumfiel

FEATURES

Top

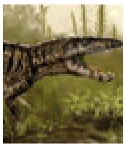
- **Research at Janelia: Life on the farm**



Five years in, has a lofty experiment in interdisciplinary research paid off?

- M. Mitchell Waldrop

- **Dark days of the Triassic: Lost world**



Did a giant impact 200 million years ago trigger a mass extinction and pave the way for the dinosaurs?

- Roff Smith

COMMENT

- **Climate policy: Letting go of Kyoto**

A preoccupation with binding commitments blocks progress in climate-change negotiations. It is time to correct course, says Elliot Diringer.

- **Environmental science: Good governance for geoengineering**

  Phil Macnaghten and Richard Owen describe the first attempt to govern a climate-engineering research project.

BOOKS AND ARTS

Top

- **Neuroscience: Neanderthals in mind**

  Clive Gamble relishes the inside story on the cognitive abilities of our fossil relatives.

  - Review of *How To Think Like a Neandertal*

    Thomas Wynn & Frederick L. Coolidge
- **Books in brief**
- **Cosmology: A life in space-time**

  George Ellis appreciates a Stephen Hawking biography that highlights the epochs of an illustrious career — and the personality behind them.

  - Review of *Stephen Hawking: An Unfettered Mind/His Life and Work*

    Kitty Ferguson
- **Environment: In at the deep end**

  A Dublin exhibition inspires a practical approach to water sustainability, finds Anthony King.

  - Review of *Surface Tension: The Future of Water*

CORRESPONDENCE

Top

- **Antarctica: Fishery threatens protected ocean**
  - Amélie Lescroël & David Grémillet
- **Women: Sexist fiction is alienating**
  - Ylaine Gerardin & Tami Lieberman
- **Women: Latent bias harms careers**
  - Pieter van Dokkum
- **Physical sciences: Research council will support excellence**
  - David Delpy & John Armitt
- **Agriculture: Risk assessment for Brazil's GM bean**
  - Rubens Onofre Nodari

OBITUARY

Top

- **Herbert Hauptman *(1917–2011)***

  Mathematician whose theories reveal the shapes of molecules from scattered X-rays.

  - Carmelo Giacovazzo

CAREERS

FEATURES

Top

- **A roll of the dice**

  For some, a lack of tenure creates a dynamic lab environment. For others, it's a gamble not worth taking.

  - Karen Kaplan

CAREER BRIEFS

Top

- **Movement in the ranks**

  US institutions dominate top academic ranks, but China is rising fast.
- **Mentors wanted**

  Mentoring service seeks more recruits.
- **Better student stability**

  Change in PhD student status would draw more researchers to Europe, says Eurodoc.

naturejobs job listings and advertising features

FUTURES

- **The loneliness of the long-distance panda**

  Bear necessities.

  - Jacey Bedford

RESEARCH

NEWS & VIEWS

INSIGHT: SILICON ELECTRONICS AND BEYOND

ARTICLES

Top

LETTERS

Top

- Cornelia Schönbauer, Jutta Distler, Nina Jährling, Martin Radolf, Hans-Ulrich Dodt ***+ et al***

- **A heteromeric Texas coral snake toxin targets acid-sensing ion channels to produce pain**
  - Christopher J. Bohlen, Alexander T. Chesler, Reza Sharif-Naeini, Katalin F. Medzihradszky, Sharleen Zhou ***+ et al***

  See also

  - News & Views by Olivera & Teichert

- **Homeostatic response to hypoxia is regulated by the N-end rule pathway in plants**
  - Daniel J. Gibbs, Seung Cho Lee, Nurulhikma Md Isa, Silvia Gramuglia, Takeshi Fukao
  - ***+ et al***
- **Oxygen sensing in plants is mediated by an N-end rule pathway for protein destabilization**
  - Francesco Licausi, Monika Kosmacz, Daan A. Weits, Beatrice Giuntoli, Federico M. Giorgi
  - ***+ et al***
- **Structural basis of RNA recognition and activation by innate immune receptor RIG-I**
  - Fuguo Jiang, Anand Ramanathan, Matthew T. Miller, Guo-Qing Tang, Michael Gale
  - ***+ et al***
- **Temperature-scan cryocrystallography reveals reaction intermediates in bacteriophytochrome**
  - Xiaojing Yang, Zhong Ren, Jane Kuk & Keith Moffat

THIS WEEK

EDITORIALS

CONSERVATION Sad goodbye to poached rhinos and frozen snails **p.268**

WORLD VIEW Carbon count sends rich nations down the UN rankings **p.269**

JAM TOMORROW Jays stash different snacks for greater choice **p.271**

Desirable partners

US refusal to collaborate with China on space science is short-sighted and misguided, from both a scientific and a pragmatic standpoint.

Space researcher Ji Wu is invigorating space science in China. He has a new, well-funded space programme, the ear of the government and a growing list of projects. China's launch capacity is set to triple, which means that it can take the lead in the launch of increasingly large and interesting missions.

A fluent English speaker and well connected in the United States, Wu is the first Chinese vice-president of the international Committee on Space Research, a position that brings unprecedented prominence to his country's space science (see page 276).

It is a perfect opportunity for the United States to establish deeper connections with China, and the ideal time to do so. As in other fields, the Chinese follow closely what is happening in the United States and look for opportunities to collaborate and learn. With many Chinese researchers returning to China carrying experience gained from the United States or, as in Wu's case, Europe, such collaborations will become increasingly important.

But that is not what is happening. Restrictions on the interaction of US scientists with their counterparts in China have been tightened to the point at which even having meetings could be considered illegal. Led by congressman Frank Wolf (Republican, Virginia), the United States has legislated and looks set to enforce a broad ban on collaborations with China on programmes funded by the Office of Science and Technology Policy or NASA. As recent high-level legal action demonstrates, Wolf is aiming to enforce the ban even on meetings between scientists from the two countries (see *Nature* **478,** 294–295; 2011).

Of course, it is right that the United States should be careful. There are justified fears of reverse engineering, and the country should collaborate on projects that would not put nationally sensitive technology at risk. But it should also be sensible and keep the doors of communication open. A blanket ban on sharing not just scientific missions but also ideas will hurt the United States more than China.

It can be reasonably argued that, given its dominance in space research and exploration, the United States does not currently have much to gain from collaborations with China. But that is surely a short-term situation and a dangerously short-term view. Over the next ten years there will be plenty of Chinese missions, such as the Solar Polar Orbit Radio Telescope and the KuaFu mission, both of which will study space weather, and which US scientists would love to be involved in to extend the value of their own missions. Anyway, as China showed with its successful space-docking manoeuvre this month, it can go it alone. Chinese students and space-science administrators are already following the US space science programme very closely. Through published road-map documents and scientific papers they already know, in broad brush, what is happening there.

The United States is well aware of the benefits of international collaboration in costly space science. It established NASA's Missions of Opportunity programme specifically to help its scientists use a small budget to latch on to bigger projects run by other countries — to be, as NASA states, "part of a non-NASA space mission of any size and having a total NASA cost of under $35 million". As budget restrictions bite, this activity will become ever more important. And the opportunities will increasingly come from China. European scientists are already building the bridges, as they did for China's Double Star mission to study Earth's magnetosphere, to move such collaborations forward.

"The United States should keep the doors of communication open."

In the congressional hearings on the collaboration ban, Wolf said that the United States had "no business" cooperating with China to help it develop its space programme. "China is taking a more assertive posture globally, and their interests rarely intersect with ours," he said. Such cold-war-era language is unhelpful. The perspective is worse.

By meeting with and, under strict conditions, collaborating on missions with Chinese space scientists, the United States will be able to build potentially beneficial diplomatic relations at the same time as keeping a close eye on Chinese space technology. Ultimately, in such collaborations the United States would be helping itself much more than it would help China. ■

Climate change

Negotiations in Durban over greenhouse-gas emissions should not try to revive Kyoto.

In a memorable scene in Al Gore's film on global warming, *An Inconvenient Truth*, the former US vice-president lampoons a cartoon of a pair of scales that weighs the Earth against a stack of gold bars. Gore's point is that any attempt to compare the merits of the two is ludicrous given their relative importance in the grand scheme of things. It would be easy to satirize reports that the organizers of next year's Rio+20 Earth Summit in Brazil are considering a two-week postponement to avoid a clash with celebrations for the Diamond Jubilee of Queen Elizabeth II in the United Kingdom, which they fear will hold more appeal for politicians, particularly those from Commonwealth countries. Easy — but not necessarily wrong. If the world is to address the myriad environmental problems that scientists have identified, then at some point it will have to give them the attention and the priority they deserve. (And that comes from a journal with its headquarters just a few miles from Buckingham Palace — sorry, Ma'am.)

A good place to start would be the international negotiations on global warming that reopen in Durban, South Africa, later this month. If optimists were right to herald the tentative steps made last year in

Mexico as a new dawn following the chaos of the 2009 Copenhagen meeting, then the Durban negotiations must now make a break from the past. Already, familiar battle lines have been drawn, and flags flown on stand-offs such as the future of the Kyoto Protocol, the global agreement that sets targets for emissions reductions. For years, environmental campaigners at the United Nations climate summits would stalk the corridors and the press room and rapidly correct anyone who claimed that the Kyoto Protocol expired in 2012. It was only the first phase of the agreement that would end, they insisted, finding hope in the implicit promise that other phases would follow. No longer — one of the hottest debates at Durban will probably boil down to whether the protocol will continue in its present form at all.

In a Comment on page 291, Elliot Diringer of the Center for Climate and Energy Solutions in Arlington, Virginia (formerly the Pew Center on Global Climate Change) makes the case that it should not. His argument — that the protocol has become an obstacle to international progress and should be consigned to history — will be popular along the interstate in Washington DC.

It is certainly pragmatic: the odds of China and the United States taking on binding emissions targets, for now, he says, are "nil", which will keep away Japan, Canada and Russia, and so fatally punch a hole below the waterline in the common-but-differentiated approach taken under Kyoto. "A binding-or-nothing mentality," has underpinned the climate talks for too long, Diringer says. "And the result often has been nothing."

Advocates of the multibillion-dollar carbon market established across Europe as a direct result of the Kyoto agreement would no doubt disagree with that assessment — as would the US airlines fighting tooth and nail to avoid being dragged into the emissions-trading scheme from next year. Many developing countries, too, would defend Kyoto, if only because it has made no serious demands of them and they enjoy seeing their wealthier rivals squirm.

But the world has changed since the formative years of the Kyoto Protocol in the 1990s, when it neatly allocated its countries into two camps — rich and poor — divided by a common purpose.

As Diringer points out, some 58% of global emissions now come from developing countries, and although a handful of rich nations still bear a heavy historical burden for global warming, it is unrealistic to expect today's politicians, who can barely look forward more than the next four or five years, to look back two centuries into the past.

"Dogmatic adherence to the protocol is now a political liability."

One of the goals of Kyoto was to make a relatively small dent in emissions, with the prospect of significantly bigger dents to come. Without the world's two largest polluters — the United States and China — on board that now seems impossible. Another goal was to establish and test an international architecture for reducing greenhouse-gas emissions and eventually scale it up. Without the world's two largest polluters, that now seems pointless.

To ditch the agreement — the only global regulation on greenhouse gases — may seem a dramatic move, and in a way it is, particularly for those who have long believed in it. But the implications need not be severe. Europe can, and should, maintain its carbon market and its commitments, just as the offset mechanism developed under the protocol can continue. The real benefits of Kyoto — practical experience and institutional structures — can endure without it.

Like it or not, a dogmatic adherence to the protocol is now a political liability that threatens cooperative action (however limited) over climate change — such as deals to secure finance for the most affected countries to help them with strategies for adaptation. There is no need to kill it. The treaty is already weakened and will prove hard to revive. The Durban meeting should be where the Kyoto Protocol, as we know it, goes to die. ■

Ex factor

Demise of snails in a New Zealand freezer is a sign of the times.

The world just got a little smaller. If you go down to the woods today in search of a western black rhinoceros (*Diceros bicornis longipes*) you'll be out of luck. Those at the International Union for Conservation of Nature, whose task it is to maintain the Red List of endangered species, have been looking high and low for western black rhinos for some time, but in vain. Last week, they called off the search and declared it extinct. Most of us will never have knowingly met a western black rhino. One feels a keen sense of its passing nonetheless — a sensation to which we are becoming accustomed. Rhinophiles will also, no doubt, be aware that the northern white rhino (*Ceratotherium simum cottoni*), a cousin of the late western black rhino, is on the brink of extinction, and that the last Javan rhino (*Rhinoceros sondaicus*) outside Java is also believed to have disappeared.

Conservation news is not all bad. The efforts of conservationists to rescue populations from the wild, breed them in captivity and reintroduce them, sometimes pay off. Przewalski's horse (*Equus ferus przewalskii*) was listed as extinct in the wild in 1996, but was brought back after a captive breeding programme, and the wild population is now believed to exceed 300. The Arabian oryx (*Oryx leucoryx*) is also on the up, albeit gingerly. Thanks to captive breeding and reintroductions — intentional or otherwise — the howls of wolves are heard once more where they had been absent for centuries, and wild boar (*Sus scrofa*) are believed to infest parts of Britain with a vigour that would shame an urban cockroach.

Less well publicized, perhaps, are the woes of endangered creatures too small, obscure or superficially revolting to attract headlines. Conservationists have long understood that the public categorizes creatures into two kinds — Cute and Yucky. The land snail *Powelliphanta augusta* tends, arguably, to fall into the latter class. Mature individuals grow to the size of a fist. A rapacious carnivore, it survives by sucking worms out of the ground. Clearly, it is a species that only its mother could truly love. It was discovered in 1996 in a remote mountain ridge on the South Island of New Zealand, its sole known place of residence. Unluckily for the marauding mollusc, its entire range was due to be demolished to make way for an opencast coalmine. About 4,000 snails were caught and released in another part of the area, with 1,600 being placed at their preferred temperature of 10 °C in chiller units in a government conservation-department facility. Unfortunately, a fault in a sensor plunged the temperature in one of the units to zero, and 800 of the snails — a sizeable fraction of the entire species — froze to death. The fault was not noticed immediately because it happened over a public holiday.

The incident highlights an important fact long known in conservation biology, that as species shrink in number, they become ever more vulnerable to sudden mishaps. To suffer because of, say, an avalanche or a brushfire is unfortunate — after all, species have evolved and become extinct innumerable times throughout Earth's history without the interference of *Homo sapiens*. So it is sad that *P. augusta* has come closer to extinction as a result of people's efforts to prevent such an eventuality. But one might, if one were so minded, also look askance at the decision to plonk an opencast mine on the snail's habitat. How different it might have been had the snails been able to disguise themselves as fluffy polar bear cubs or baby pandas. Conservationists can only do so much. When backed with political will, they can do much more. ■

➲ NATURE.COM
To comment online, click on Editorials at: go.nature.com/xhunqv

Time to stop celebrating the polluters

The United Nations must include sustainability in its quality-of-life index to encourage countries to develop responsibly, says **Chuluun Togtokh**.

The United Nations Development Programme this month released its annual league table of countries judged according to their state of development. Who leads this ranking? The usual suspects: the United States, Canada and Australia are all among the top six. My own nation, Mongolia, languishes in 110th place.

The UN goes out of its way to promote sustainable development, yet the Human Development Index (HDI) mostly ignores sustainability. Worse still, the index celebrates gas-guzzling developed nations. It is time that this failure — hidden in plain sight — was exposed and corrected.

The HDI has set straightforward benchmarks for countries and international organizations for more than 20 years. Its success and influence owes much to its simplicity. The index brilliantly summarizes development and quality of life in a given country using health, education and income levels. Yet it fails to cover an increasingly crucial question: how responsible is that development? With Earth's human population reaching 7 billion in the past month, it is reasonable to question the UN's true commitment to sustainability.

In the current HDI, developed nations and oil-rich countries are placed highly without regard to how much their development paths cost the planet and imperil humanity's future development. There is an assumption that natural resources are unlimited, and little regard is given to the fundamental changes to Earth's biological, physical and chemical processes that result from development. Either we have unbridled optimism that a miracle will occur, or our scepticism about our ability to overcome this massive challenge is so paralysing that we do not even bother to try.

PROGRESS IN THE **DEVELOPMENT INDEX** HAS COME AT THE COST OF **GLOBAL WARMING.**

In 1992, the first UN Earth Summit in Rio de Janeiro, Brazil, defined the three pillars of sustainable development: economic, social and environmental growth. Globally, humanity has had remarkable success with the first two of these. But we have failed to tackle all three dimensions simultaneously, owing to reductionism, fragmentation, division and territoriality. The HDI is emblematic of this fragmented approach.

As the UN prepares to return to Rio de Janeiro for the Earth Summit 2012, it must lead by example. From next year, it should change the way it calculates the HDI. The revised index should include each nation's per capita carbon emissions, and so become a Human Sustainable Development Index (HSDI).

Per capita emissions are a simple, available and quantifiable indicator, and this month's report announcing the HDI did include some important analysis of them. Emissions are positively and strongly correlated with income; less so with the HDI; and not at all with health and education. And in general, the faster a country's HDI improves, the faster its carbon dioxide emissions increase. The bottom line is that progress in the HDI has come at the cost of global warming. But these environmental costs are related only to economic growth, not to broader gains in the HDI, and the relationship is not fixed. Some countries have advanced in both the HDI and environmental sustainability.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/g6tlgh

How would inclusion of emissions affect the HDI? To find out, I recalculated the index using the UN's published methodology, but taking per capita emissions into account. The resulting HSDI gives some interesting results.

Australia, the United States and Canada fall straight out of the top 10: Australia slides from 2nd place to 26th, the United States drops from 4th to 28th, and Canada falls from 6th to 24th.

Cultures that value moderation do well in this sustainability index: Norway remains in the top position, Sweden rises from 10th to 2nd and Switzerland moves from 11th to 3rd. But anyone who has visited the Nordic countries will recognize that moderation need not compromise a high standard of living. And for the first time, an Asian state appears in the top ten. Hong Kong rises from 13th place to 4th. Japan and South Korea, originally just outside the top ten, move down by only one or two places.

Noticeably, oil-producing countries and those with intensive oil use drop the most. The United Arab Emirates, Brunei Darussalam, Qatar, Luxembourg and Bahrain are no longer listed in the 'Very High Human Development' quartile.

Using the HSDI, Mongolia advances slightly. My country is likely to become one of the fastest growing economies in the world, but the current HDI offers no encouragement for it to grow sustainably. Ulaanbaatar is already one of the worst capital cities in the world for air pollution. The country's water, forage and forest resources are depleted. Mongolia is at a turning point in environmental, social, economic, political and cultural development. We urgently need international collaborations to preserve our natural and cultural systems and introduce green technologies.

It seems part of human nature at all levels to compete, and this can be harnessed. The HDI has shifted the target of development beyond the almighty dollar; the proposed HSDI would go one step further, and change the role models for development. We need such a change because, if the UN continues to encourage countries such as Mongolia to aspire to the US lifestyle, we will all be in serious trouble. ■

Chuluun Togtokh *is a professor of ecosystem and sustainability sciences at the National University of Mongolia in Ulaanbaatar and vice-chair of Mongolia's Global Change National Committee.*
e-mail: chuluun@warnercnr.colostate.edu

SEVEN DAYS *The news in brief*

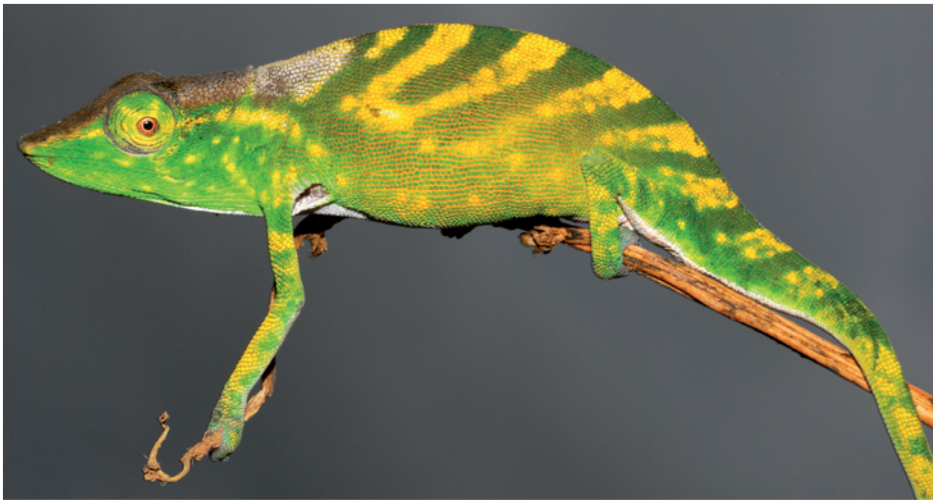
POLICY

AIDS-free goal

Achieving a generation worldwide without AIDS is now a policy priority for the United States, because it has become a feasible goal, Secretary of State Hillary Clinton announced on 8 November at the National Institutes of Health in Bethesda, Maryland. The goal could be reached by combining three proven strategies, she said: prevent mother-to-child transmission, increase adult male circumcision rates, and scale up anti-retroviral treatment for people with HIV/AIDS. See go.nature.com/fdd1f4 for more.

Personhood falters

Voters in Mississippi have rejected a state constitutional amendment to redefine 'person' as "every human being from the moment of fertilization, cloning, or the equivalent thereof". Had the initiative passed on 8 November, it would have outlawed abortion in the socially conservative state and could have restricted the use of reproductive technology and certain contraceptives. The amendment earned 42% of the vote — a significant increase over the 27% and 29% achieved by previous attempts to pass similar amendments in Colorado. The 'personhood' movement is set to try its luck in elections in several other states in 2012 (see *Nature* **479,** 13–14; 2011).



J. KÖHLER

Red List reptiles

Tarzan's chameleon (*Calumma tarzan*, pictured), was discovered only last year in forests in Madagascar, but it has almost immediately been classed as critically endangered, following its assessment for the International Union for Conservation of Nature's Red List. The lizard, which is fast losing habitat because of slash-and-burn agriculture and logging, joins 21 other Madagascan reptile species in the critically endangered category. See Editorial, page 268 for more on this year's Red List.

Polar-bear concern

Canada's environment ministry has declared the polar bear to be a "species of special concern". The 10 November decision means that a management plan to protect the species must be produced within three years. Around two-thirds of the world's polar bears live in Canada. The Center for Biological Diversity, an environmental campaign group based in Tucson, Arizona, complained that the bear should have been listed as 'threatened' or 'endangered', which would have prohibited some types of hunting and established a protected 'critical habitat'.

Nuclear Iran

Iran has pursued work related to the development of a nuclear weapon, the International Atomic Energy Agency (IAEA) said in a report released on 8 November. The assessment marks a significant departure from the agency's previous reports, which have previously criticized Iran for hiding nuclear facilities. See page 282 for more.

Pipeline postponed

The US government has delayed a key decision on whether to approve a 2,700-kilometre pipeline to carry oil from tar sands in Alberta, Canada, to the coast of Texas. On 10 November, the state department said that it needed to explore alternative routes through Nebraska for the Keystone XL pipeline, after protests from landowners in the state and from environmental groups who oppose the use of tar-sands oil. The delay means that there can be no decision on the project until after the 2012 elections in the United States.

RESEARCH

NSF goes boldly

In an effort to encourage "bold" interdisciplinary research, the US National Science Foundation (NSF) is starting a grants programme that will skip the agency's normal process of external peer review. The programme, named Creative Research Awards for Transformative Interdisciplinary Ventures (CREATIV), will accept only high-risk proposals that cross traditional research disciplines. Requests can be

JOURNALISM FELLOWSHIP

Canada's International Development Research Centre and *Nature* are offering a six-month, full-time, fully funded science journalism fellowship to an English-speaking Canadian citizen or permanent resident of Canada. See go.nature.com/ntrtp4 for details on how to apply.

for up to US$1 million over a period of up to five years, and will be subject only to internal review from the NSF.

PEOPLE

RNA pioneer dies

Har Gobind Khorana, the biochemist who rose from humble origins in rural India to win the Nobel Prize in Physiology or Medicine in 1968, died on 9 November aged 89. He won the prize while working at the university, for discovering how RNA codes for the synthesis of proteins. Khorana shared the award with Robert Holley and Marshall Nirenberg.

Stapel returns PhD

Diederik Stapel (**pictured**), the prominent Dutch psychologist who was found to have faked data in at least 30 research papers, has voluntarily surrendered the PhD he earned from the University of Amsterdam in 1997, the university announced on 10 November. Stapel declared in a statement that his recent behaviour did not befit the holder of a doctorate. A 31 October report revealing Stapel's misconduct (see *Nature* **479,** 15; 2011) said it was impossible to determine if his PhD was fraudulent, and recommended that the university investigate

P. VAN EIJNDHOVEN

whether the degree could be withdrawn on grounds of unworthy conduct.

Turkish revolt

Almost half of the members of the Turkish Academy of Sciences resigned last week, protesting against government decrees that threatened the academy's autonomy. By 11 November, 64 members (46%) had resigned, after a 4 November statute did not significantly soften a 27 August decree that gave the government power to appoint the president and most members of the academy, and enabled it to nominate top personnel in research-funding agency TÜBİTAK.

DOE departure

Steven Koonin, the undersecretary for science at the US Department of Energy, is leaving the agency after two and a half years. The department's head, Steven Chu, told employees on 8 November that Koonin would leave on 18 November to join the Science and Technology Policy Institute in Washington DC. Koonin, previously chief scientist at energy firm BP and provost of the California Institute of Technology in Pasadena, was one of Chu's highest-profile appointments. See go.nature.com/ktaxby for more.

BUSINESS

Depression setback

Shares in US biotechnology firm Targacept of Winston-Salem, North Carolina, plunged by more than 50% after it announced on 8 November that an experimental antidepressant drug had failed in the first of four late-stage clinical trials. The drug, TC-5214, is the first in a new class of depression treatments that work on the brain's nicotine receptors. See page 278 for more.

Stem-cell departure

Geron, the first company to gain US approval for a clinical trial using human embryonic stem cells, is to walk away from the scientific field that it helped to create. The firm, based in Menlo Park, California, said on 14 November that it will stop its stem-cell programme and instead focus on cancer therapies. Among the casualties is Geron's landmark stem-cell treatment for spinal cord injuries. See go.nature.com/n6bbyu for more.

EVENTS

Soyuz success

A Soyuz rocket successfully launched from the Baikonur Cosmodrome in Kazakhstan this week carrying three astronauts to the International Space Station (ISS). Anton Shkaplerov, Anatoly Ivanishin and Dan Burbank were the first people to fly on a Soyuz since a cargo craft atop a Soyuz-U rocket crashed in August. If the 13 November launch had not gone ahead, the ISS might have had to be left temporarily uncrewed, as the three astronauts now in residence are due to leave the station on 21 November.

Mars no go

As *Nature* went to press, the Russian space agency Roscosmos had been unable to re-establish contact with its Phobos-Grunt probe. The soil-return mission was launched on 9 November, but became stuck in an Earth orbit. See page 275 for more.

COMING UP

17–19 NOVEMBER
The fifth biennial World Science Forum in Budapest discusses 'The Changing Landscape of Science'.
www.sciforum.hu

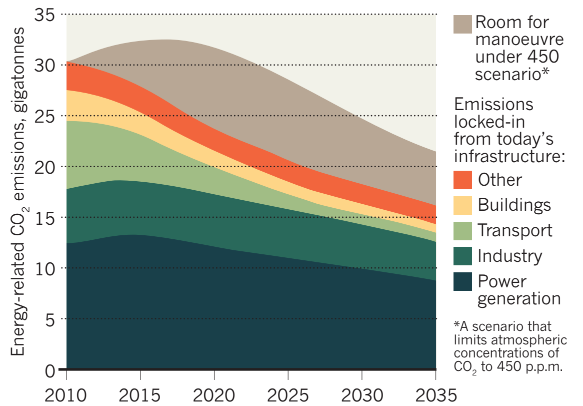
23 NOVEMBER
A US bipartisan congressional committee reaches its deadline for finding ways to cut at least US$1.2 trillion from the country's deficit over the next ten years. Its formulae will be keenly watched by science lobbyists.

TREND WATCH

If significant policy action is not taken, it will be impossible to hold atmospheric carbon dioxide concentrations below 450 parts per million, the target thought to give an even chance of limiting global warming to 2 °C. According to *World Energy Outlook 2011*, released on 9 November by the International Energy Agency, assuming no new policies are implemented, new infrastructure to sate the world's energy needs will mean that exceeding the target level cannot be avoided after 2017.

APPROACHING A CARBON LIMIT

Existing infrastructure will emit 80% of the CO_2 emissions allowed by a scenario that holds concentrations below 450 p.p.m.



SOURCE: IEA

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

VICTOR O LESHYK

RIA NOVOSTI

Engineers inspect the Phobos-Grunt probe in happier times.

PLANETARY SCIENCE

Russia gets the red planet blues

Phobos probe failure puts planetary comeback in doubt.

BY ERIC HAND

It was the largest planetary mission in the history of space exploration, bearing Russia's hopes of recapturing Soviet-era glory in Solar System exploration. But instead of rocketing off on a mission to return soil from the Martian moon Phobos, Phobos-Grunt is stuck in Earth orbit. Barring a miraculous restarting of its engines, it will make a fiery fall to Earth, probably by the year's end.

The Russian planetary programme could plunge along with it. Roscosmos, the Russian space agency, had hoped that a successful robotic surface-sample return — a feat that NASA has not yet achieved — would, in one stroke, erase the demons of past failures. Instead, the agency finds itself exactly where it stood 15 years ago after the launch of the ill-fated Mars 96 mission, a previous massive assault on the red planet: unable to leave Earth orbit. "It was over-ambitious," says Roald Sagdeev, former head of the Space Research Institute (IKI) in Moscow, who is now at the University of Maryland in College Park. He sees a suite of future planned missions now in jeopardy, as well as chances for international cooperation. "Opponents will say: why should we waste money on planetary missions?"

Engineers at Roscosmos have been trying daily to communicate with the stalled spacecraft, which at 13,500 kilograms is more than twice as massive as the US–European Cassini mission to Saturn, the largest probe sent beyond lunar orbit so far. As Phobos-Grunt (literally Phobos-soil) seems to be undamaged and its fuel tanks full, Roscosmos hoped to reboot the craft, restart its engines and send it on its way to a rendezvous with Mars' largest moon in early 2013. But all communication attempts have failed, says Alexander Zakharov, the mission's project scientist at the IKI, which manages the 20 instruments on the scientific payload. The window for getting the spacecraft to Mars — determined by the orbital paths of Earth and Mars — will shut on 21 November. "We have a few days more," says Zakharov.

After launching on 8 November from Baikonur Cosmodrome in Kazahkstan, the first stage of the Zenit rocket worked perfectly. But amateur skywatchers in South America, who were enlisted by Russia to supplement its meagre tracking resources, didn't see the expected burn of the second stage, says Ted Molczan, an amateur observer in Toronto, Canada, who runs a satellite-observing group online called SeeSat-L.

A fault in the second-stage control systems seems to be the immediate cause. But journalist Anatoly Zak, the founding editor of russianspaceweb.com, says that the deeper problems are institutional. A culture of lax testing meant that some problems lay undetected, whereas known problems were ignored at the higher levels, he says. "The project was doomed from the beginning."

Sagdeev adds that the Russian programme's strength was always in hardware, not control systems and software. "This is the legacy of the Soviet space programme," he says. "The emphasis was on building rockets." ▶

Friction with the upper atmosphere is steadily dragging on the orbiting spacecraft, now just over 200 kilometres above Earth at its lowest point. Zakharov says it is expected to fall in mid-December. Molczan, who has performed his own analysis of the orbital decay, says it could be as late as January.

Re-entry could make for a massive fireball, as most of the weight of the spacecraft is in its liquid fuel — much of it meant for a return trip from Mars. A statement on the Roscosmos website says that the craft will blow up in the heat of re-entry, which would reduce the risk of anything reaching the ground. But Jonathan McDowell, an astronomer and satellite watcher at the Harvard-Smithsonian Observatory in Cambridge, Massachusetts, says that it's also possible that the fuel, made up of hydrazine and dinitrogen tetroxide, will freeze in space during the spacecraft's initial descent, increasing the chance that some of the craft and its contents might survive the plunge. "Now you have a lump of a couple of tonnes of toxic sludge that's falling out of the sky," he says.

Even without the public-relations disaster of a crash to Earth, the damage to Russia's future planetary plans will be considerable. Zakharov says it will be difficult for Roscosmos to consider repeating the US$163-million Phobos mission. The agency and its primary contractor, Lavochkin, had also wanted to pursue missions to the Moon, Mars, Venus, Mercury and even Europa, the icy moon of Jupiter. But the plans, never firm, are now more uncertain. A 2014 launch of the Luna-Resource mission, a Moon lander paired with an Indian-built orbiter and rover, is still on track, Zakharov says. But after the loss of Phobos-Grunt, which carried a small Chinese satellite, India will be rightly concerned, says Sagdeev. "If something similar happens to the Indian payload, it would be a real disaster."

"The project was doomed from the beginning."

In the old days, failures could be tolerated, because the Soviet Union would launch mission after mission until each problem was fixed, says Louis Friedman, former director of the Planetary Society in Pasadena, California, and principal investigator of a small astrobiology experiment on Phobos-Grunt. But with no money for frequent launches, Roscosmos and Lavochkin — which has designed and built all Russia's interplanetary missions since the dawn of the space age — will have to change their cultures, Friedman says. "The hope is that there will be a real shake-up," he says. "Do I think that will happen? History says it won't." ■

CHINAFOTOPRESS/PHOTOSHOT

The Shenzhou-8 spacecraft (above, before launch) has docked successfully with the Tiangong-1 module.

SPACE EXPLORATION

China forges ahead in space

Mars-probe problems are a minor blip in a bold strategy.

BY DAVID CYRANOSKI

The likely demise of Russia's Phobos-Grunt mission has dashed China's hopes for its first Mars orbiter, Yinghuo-1, which was piggybacking on the larger craft (see page 275). But it is a relatively small setback for a nation that has notched up a string of high-profile space successes in recent years, including this month's 'heavenly kiss' of two unmanned orbiters, Shenzhou-8 and Tiangong-1 — a milestone in the country's effort to build a manned space station, the Tiangong ('Heavenly Palace'), by the end of the decade.

And ambitious moves on the ground suggest that China will increasingly be able to develop and launch its probes without partnering with other nations. In July, the Chinese Academy of Sciences (CAS) in Beijing opened its National Space Science Center (NSSC), which will take charge of overall planning for the country's space science. "China was a space country without a space science programme," says Ji Wu, the centre's director. Now that the CAS "has got government support to manage space science missions as a series", he says, "it will lead to a new era for space science in China".

For years, the lack of a clear national strategic plan that prioritized missions has hampered researchers' efforts to launch space telescopes or planetary probes. The convoluted relationships between the various Chinese agencies involved in funding, building and launching satellites impeded many mission proposals, and decisions about their fate were at best "opaque", according to one European researcher who has collaborated with Chinese space scientists.

Centralized planning under the NSSC could change that. The centre already has about 450 staff, including 50 scientists, inherited from its predecessor, the Center for Space Science and Applied Research, which last year had a budget of 300 million renminbi (US$47 million). (Individual missions have separate funding.) That will grow to 700 million renminbi as the centre develops space science missions over the next few years.

First in the queue is an orbiting X-ray observatory, HXMT, with a budget of about 900 million renminbi and due for launch in 2014. Then there is the KuaFu mission, which aims to reach space the following year, to study the Sun's effects on space weather. Later missions could fly on China's Long March 5 rocket, expected to enter service in 2014. Its ability to loft payloads of 14 tonnes into a highly elliptical orbit should allow China to launch interplanetary missions as large as Phobos-Grunt.

"In twenty years, China will be dominating the United States."

The NSSC's mission selection and planning procedure is "more or less the way we prepare missions at ESA", says Philippe Escoubet, a mission manager at the European Space Agency's (ESA's) Science and Robotic

Exploration Directorate in Noordwijk, the Netherlands. "It's very much welcome."

Escoubet collaborated with Chinese scientists on the Double Star mission, a joint effort with ESA that launched two satellites, in 2003 and 2004, to study Earth's magnetosphere. He believes that, as China devotes more resources to space science, there will be greater opportunities for collaboration.

CAS president Bai Chunli, in his speech at the NSSC inauguration, confirmed that the centre will endeavour to "deepen international cooperation", adding that "scientists from around the world will be able to access data" from the missions.

But Wu must win the confidence of potential collaborators who have been frustrated in the past by bureaucracy and delays when they tried to work with China. Indeed, several scientists interviewed by *Nature* for this article were unwilling to speak on the record, fearing that future partnerships with China could be jeopardized.

Some worry that the new missions, like China's space station, may be exclusively national projects. "In principle, it is relatively advantageous to European researchers to be involved in Chinese missions," says one UK space scientist. "However, there has not yet been a Chinese mission with an instrument selection process that was open to the wider world, as opposed to 'by invitation.'"

Collaboration with scientists in the United States is certainly unlikely. That country has always vetoed Chinese participation in the International Space Station, and congressional antipathy towards China was ramped up earlier this year. In April, congressman Frank Wolf (Republican, Virginia), who chairs the subcommittee that funds NASA, modified a spending bill to prevent the agency from using federal funds on joint projects with China. But if China's planned space science missions go ahead "and the United States is locked out, it will be a huge missed opportunity", says a US-based astrophysicist familiar with Chinese science.

"China is firmly committed to upping its profile in space science, both to expand its technological base and to boost its domestic and international prestige," says space security expert Clay Moltz of the Naval Postgraduate School in Monterey, California. "As long as China's economy continues to deliver strong growth, I expect space science missions will continue to expand."

A European scientist who has worked with China's space scientists agrees that its space science will flourish, particularly as Chinese senior scientists return from research positions in the West. "In twenty years, China will be dominating the United States," he says. "There is a sense of national urgency and dedication in China, and the rate at which they learn is phenomenal. It won't take that long." ■ SEE EDITORIAL P.267

ENVIRONMENT

Gulf ecology hit by coastal development

Dubai's artificial islands are affecting marine ecosystems.

BY DANIEL CRESSEY

Untrammelled development, weak regulatory oversight and a lack of scientific monitoring are seriously threatening ecosystems along the coast of the Gulf, according to an extensive assessment of the region's marine environment.

Sea-front projects ranging from desalination plants to artificial islands in the gulf between the Arabian Peninsula and Iran have transformed the entire coastline in the past few decades. More than 40% of some countries' shores are now developed. The change is happening more quickly, and with greater environmental impact, than in any other coastal region. "Things are being put in place so quickly we don't know what is going to happen," says Peter Sale, a marine ecologist at the United Nations University (UNU) Institute for Water, Environment and Health in Hamilton, Ontario, Canada, who co-authored the report[1].

The report synthesizes existing research with the UNU team's own assessments of the impact of projects such as Palm Jumeirah, an artificial archipelago more than 5 kilometres wide in Dubai in the United Arab Emirates. Some 94 million cubic metres of sediment were dredged up to make the islands. Sale says that such projects are "so substantial that they have changed the ecology in ways that are only going to become clear in decades".

➲ NATURE.COM
Read more from the region at Nature Middle East:
go.nature.com/9aofs6

Water around some parts of the islands can remain almost stationary for several weeks, increasing the risk of algal blooms. And although fish have colonized the new environment, they are not all the same species that were there before.

The report says that legislation and regulatory frameworks are often inadequate to properly monitor such projects. Environmental assessments that could take a year in Western nations might be done far more superficially in ten weeks in the Gulf, for example. And a lack of scientific input to the development process means that there is often little or no environmental monitoring. "There is a clear deficit, not only in the capacity but in scientific knowledge, and a limited amount of scientific data on which to base decisions," says report co-author Hanneke Van Lavieren, a programme officer for coastal ecosystems at the UNU.

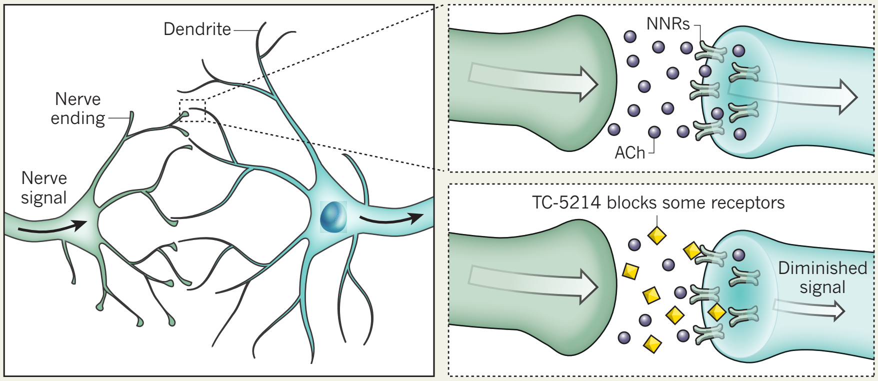
The problem is becoming increasingly urgent. The region has already lost 70% of its coral reefs since 2001, with most of the remaining reefs threatened or degraded, for example[2]. Construction of Dubai's Palm Jebel Ali, an even larger artificial archipelago, has already destroyed 8 square kilometres of natural reef, the report notes.

"All the ecological trajectories are downhill," says Charles Sheppard, an ecologist at the University of Warwick, UK, who has studied environmental change in the Gulf. "The prognosis isn't good unless there is major change." ■

1. Van Lavieren, H. *et al. Managing the Growing Impacts of Development on Fragile Coastal and Marine Ecosystems: Lessons from the Gulf* (UNU, 2011).
2. Wilkinson, C. (ed.) *Status of Coral Reefs of the World: 2008* (GCRMN, 2008).

Problems are brewing in the still waters around the artificial Palm Jumeirah archipelago.

MIXED SIGNALS
The depression drug TC-5214 binds to nicotinic receptors (NNRs) on dendrites, diminishing the effect of the neurotransmitter acetylcholine (ACh), which has been linked to depression.



CLINICAL TRIALS

Depression drug disappoints

Failure of a promising compound casts a shadow on others.

BY HEIDI LEDFORD

It would not be the first psychiatric drug to run aground in a large study after sailing through early trials. But even though TC-5214 has failed to significantly relieve major depression in a phase III trial and investors are fleeing, some analysts and scientists argue that the setback need not spell the end for the drug, nor for other compounds that act on nicotinic receptors in nerve cells.

On 8 November, Targacept, a drug company based in Winston-Salem, North Carolina, announced that TC-5214 had performed no better than placebo in one of four phase III trials. The results are a disappointment to clinicians eager for an innovative antidepressant. Because the drug exploits a previously untried mechanism, it might have helped the roughly one-third of people with depression who do not respond to current therapies[1]. "We really need new options," says Noah Philip, a psychiatrist at Brown University in Providence, Rhode Island. "People were very eager to see what this drug would do."

Results from the other trials are expected by early 2012, but some analysts are pessimistic; Targacept's stock fell by 60% after the announcement. "I was stunned by the negative outcome," says Alan Carr, an analyst for the Needham & Co investment bank in New York. "I don't have high expectations for the remaining three trials."

TC-5214 is a form of mecamylamine, a blood-pressure drug introduced in the 1950s. It targets nicotinic $\alpha 4\beta 2$ receptors (see 'Mixed signals'), which normally receive chemical signals from the neurotransmitter acetylcholine. Because excess acetylcholine has been linked to major depression[2], blocking these signals might relieve the condition.

Interest in the drug's use in neuropsychiatric disorders began in the 1990s, when nicotine was seen to reduce the symptoms of Tourette's syndrome. Searching for a similar effect with a less harmful substance, researchers at the University of South Florida in Tampa investigated mecamylamine. It did not ease the tics characteristic of Tourette's syndrome, but did relieve depression and improve mood regulation in some people[3].

The result caught the attention of Targacept, which licensed mecamylamine patents held by the University of South Florida. In 2009, the London-based pharmaceutical company AstraZeneca backed TC-5214 with US$200 million up front, and up to $1 billion in milestone payments. That year, Targacept reported that in phase II trials with an approved antidepressant, TC-5214 had few side effects and produced a six-point improvement on the Hamilton rating scale for depression.

Last week's discouraging result "is not necessarily the end of this drug, although you might think so if you just look at Targacept's share price", says Daniel Chancellor, a health-care analyst for Datamonitor, a market-research firm in London. Targacept needs to succeed in only two of its phase III trials to gain approval, says Merouane Bencherif, head of preclinical research at the company. Chancellor notes that Forest Laboratories, based in New York, reported in January that its candidate antidepressant, levomilnacipran, had failed in a phase III trial, yet in July the company announced that the drug had significantly improved symptoms in another phase III trial..

Companies such as Targacept also hold high hopes for nicotinic-receptor modulators that stimulate, rather than inhibit, the acetylcholine pathway. So far, the only approved such modulator is varenicline (Chantix/Champix), a drug to help people stop smoking that has come under fire over possible psychiatric side effects[4].

With backing from AstraZeneca, Targacept is testing two compounds against Alzheimer's disease. Other companies, including Pfizer, based in New York, have invested in nicotinic-receptor modulators to treat Alzheimer's disease and attention-deficit hyperactivity disorder (see table).

The discouraging news about TC-5214 "just shows that there are some unknowns in the biology here", says Carr, who cautions that it is too soon to write off the drug class. "I think it's just a matter of time before something does come along that makes it past that final hurdle." ■

1. Rush, A. J. *et al. Am. J. Psychiatry* **163,** 1905–1917 (2006).
2. Shytle, R. D. *et al. Mol. Psychiatry* **7,** 525–535 (2002).
3. Shytle, R. D., Silver, A. A. & Sanberg, P. R. *Biol. Psychiatry* **48,** 1028–1031 (2000).
4. *Nature* **466,** 677 (2010).

AIMING AT THE BRAIN'S NICOTINE RECEPTORS
An emerging class of drugs modulates receptors that normally bind the neurotransmitter acetylcholine.

Drug	Company	Indication	Stage
Chantix/Champix (varenicline)	Pfizer	Smoking cessation	Approved
TC-5214	AstraZeneca/Targacept	Major depressive disorder	Phase III
AZD3480	AstraZeneca/Targacept	Alzheimer's disease	Phase IIb
		Adult attention-deficit hyperactivity disorder (ADHD)	Phase II
CP-601927	Pfizer	Major depressive disorder	Phase II
AZD1446	AstraZeneca/Targacept	Alzheimer's disease	Phase I
		Adult ADHD	Phase II

FOOD SECURITY

Summit urged to clean up farming

Leading scientists say that agriculture is a 'poor relation' in global-warming negotiations.

BY NATASHA GILBERT

Delegates meeting this month in Durban, South Africa, to assess international progress on tackling climate change need to look beyond smoke stacks and car exhausts to a neglected source of emissions — agriculture.

That's the message from an international group of leading agricultural and climate scientists in a report published on 16 November. They say that agriculture is the "single largest contributor to greenhouse-gas pollution on the planet", through routes such as deforestation, rice growing and animal husbandry (see 'Farming footprint'). Emissions include nitrous oxide from fertilizer and methane from livestock, as well as carbon dioxide. With global food demand projected to double by 2050, agriculture's emissions will grow — unless farming can become dramatically more efficient. Agriculture is a "poor relation" in negotiations on strategies to mitigate climate change, says John Beddington, Britain's chief scientific adviser and chair of the Commission on Sustainable Agriculture and Climate Change, an initiative of the Consultative Group on International Agricultural Research in Washington DC, which produced the report.

The United Nations Framework Convention on Climate Change (UNFCCC), sponsor of the Durban meeting, has no specific provisions for addressing agricultural greenhouse-gas emissions. The scientists recommend that parties to the UNFCCC establish a programme to develop a global sustainable agriculture strategy, and argue that the problem deserves a larger share of international climate-change mitigation funding.

"Everyone is hoping that UNFCCC will agree to establish the agricultural work programme in Durban. If it doesn't happen we will be in a much worse position," says Tim Benton, a sustainable-agriculture researcher at the University of Leeds, UK.

One author of the report, Tekalign Mamo, Ethiopia's minister of state for agriculture and rural development, told *Nature* that policy-makers at Durban should take examples of good agricultural practice and replicate their success internationally. A successful programme in Ethiopia, for example, has given cash and food to poor households in exchange for labour on projects to improve soil quality, water supplies and infrastructure.

The report also praises Australia's Carbon Farming Initiative — the world's first national legislation aimed at reducing carbon emissions from farming and forestry, which was enacted in August. The law allows farmers and investors to generate and trade carbon credits from farming and forestry projects, and could serve as a model for similar projects in other countries.

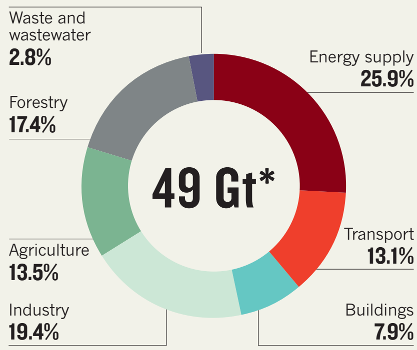
Reducing waste is a key goal: one-third of the food produced for human consumption is lost to inefficiencies in production, storage and transport, the report says.

Benton believes that the "intellectual weight" of the report's authors will help it to influence policy-makers. As well as Beddington and Mamo, they include Carlos Nobre, a climate scientist at Brazil's National Institute for Space Research in São Paulo, and Marion Guillou, president of the French National Institute for Agriculture in Paris.

Camilla Toulmin, director of the International Institute for the Environment and Development in London, hopes Benton is right. But she worries that the prospects for decisive action at Durban are poor, because governments are "distracted by the economic crisis". ■

FARMING FOOTPRINT

Greenhouse-gas emissions from forestry are largely caused by creating new farmland. When added to emissions directly from agriculture, farming is the largest source of man-made greenhouse gases.



*49 gigatonnes of carbon-dioxide equivalent per year; 2004 data

➲ NATURE.COM
Can science feed the world?
For more see:
www.nature.com/food

MEDICINE

Targeted treatment tested as potential cancer cure

Trial will deploy genetically targeted therapy early, rather than as last resort.

BY ERIKA CHECK HAYDEN

After Van VanderMeer was diagnosed with advanced lung cancer, the results of a genetic test offered some hope. Last year, the 64-year-old lawyer learned that his cancer featured a genetic rearrangement that might render it vulnerable to a drug being tested in clinical trials. But because the experimental drug, crizotinib, was being given only to patients who had failed chemotherapy, VanderMeer had to wait for more than a year to gain access to the drug. Even though VanderMeer's tumours had by then spread to both of his lungs, crizotinib vaporized them within two weeks.

VanderMeer is now doing well and hoping to continue beating the disease: more than half of patients who take the drug, made by Pfizer of New York, seem to have a better prognosis than do those who didn't receive treatment. But what if VanderMeer had started taking it sooner?

Now oncologists, pathologists and geneticists are hoping to answer that question with a study that will test whether genetically targeted treatments, applied soon enough, can cure patients of lung cancer rather than buying them a few extra months of life.

Targeted therapies have now been approved for many cancers, and it has become routine for major cancer centres to genotype patients' tumours to determine whether they might benefit from targeted drugs, in case standard treatments fail. But the clinical trial, which will be conducted by the Alliance for Clinical Trials in Oncology, a nationwide group funded by the US National Cancer Institute in Bethesda, Maryland, will test whether using targeted treatments earlier can prevent patients with lung cancer from ever reaching that point.

In the trial, tumours will be genotyped after surgery to determine whether mutations are present in a gene encoding epidermal growth factor receptor (EGFR). Mutations in this gene are targeted by many molecular therapies, including erlotinib and gefitinib, which are approved for the treatment of advanced lung cancer. Some of the patients who have *EGFR* mutations will begin taking erlotinib after surgery, instead of waiting to see whether their cancer recurs.

Although similar approaches have been tested in smaller trials, yielding mixed results, organizers say that a larger, better-defined study is needed to provide a clear answer.

"We have never tested these drugs in the right population," says oncologist Ramaswamy Govindan of Washington University in St Louis, leader of the trial. "We have never tested a group of patients who have mutations in *EGFR* and then asked the question, 'could these patients be cured by gefitinib or erlotinib?'"

He hopes to expand the analysis to include crizotinib, which targets a different genetic rearrangement and was approved by the US Food and Drug Administration in August. Other targeted therapies are in the pipeline. In a 9 November paper, for instance, a consortium of researchers from Massachusetts General Hospital and Harvard Medical School, both in Boston, and Yale University in New Haven, Connecticut, describe the results of a study that tested more than 500 patients with non-small cell lung cancer (L. V. Sequist *et al. Ann. Oncol.* http://dx.doi.org/10.1093/annonc/mdr489; 2011). The authors examined mutations in several genes relevant to therapies that have been approved or are in development (see 'Identifying targets'). Of the 353 patients with the most advanced lung cancers, 22% were matched to clinical trials appropriate for their cancer type.

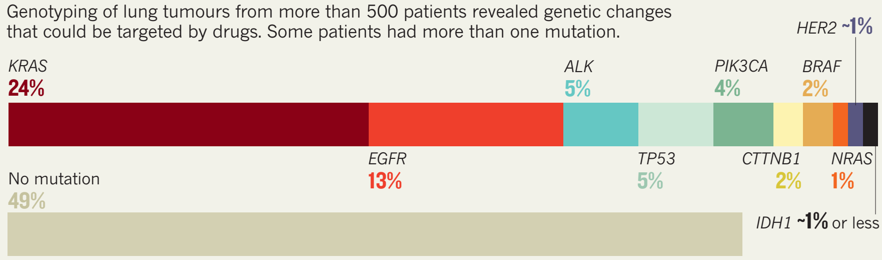
The Alliance trial will be logistically difficult. Only 10–20% of patients with non-small cell lung cancer have mutations in the *EGFR* gene; only 20% of patients are diagnosed early enough to benefit from surgery; and only a fraction of patients with the appropriate mutations will actually gain any advantage from targeted treatments. To reach their target of 400 participants, Govindan and his colleagues may need to screen as many as 1,500 people.

VanderMeer, for one, hopes that the efforts pay off — and spare other patients from what he calls the "blunderbuss" of chemotherapy.

"I'd hate for anyone to have to go through the blunderbuss before they get to the stiletto," he says. ■

IDENTIFYING TARGETS

Genotyping of lung tumours from more than 500 patients revealed genetic changes that could be targeted by drugs. Some patients had more than one mutation.

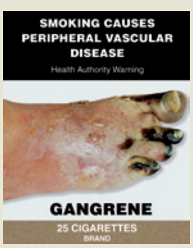


SOURCE: L. V. SEQUIST ET AL. ANN. ONCOL. (2011)

MORE ONLINE

EXPLAINER

The science behind Australia's war on tobacco advertising go.nature.com/zjrfci



AUSTRALIAN CANCER COUNCIL

MORE NEWS

- Sickle-cell mystery solved go.nature.com/dxx61h
- Ancient adaptations to parasites drove human genetic variation go.nature.com/d4es4b
- Proof found for unifying quantum principle go.nature.com/dt8syh

FROM THE BLOG

TEPCO

Japan funds Fukushima clean-up projects go.nature.com/wdfscp

NUCLEAR WEAPONS

Iran's nuclear plan revealed

Report paints detailed picture of nation's intention to build a warhead.

BY GEOFF BRUMFIEL

A report released last week by the International Atomic Energy Agency (IAEA) on Iran's alleged research into nuclear weapons assembles old intelligence into the sharpest picture yet of the weapon that Iran hopes to develop.

The 8 November report is also the IAEA's strongest statement to date that Iran's activities violate the Treaty on the Non-Proliferation of Nuclear Weapons (NPT), which explicitly prohibits the development of weapons. It "contains no new information", says Ali Vaez, director of the 'Iran Project' at the Federation of American Scientists, a think tank based in Washington DC.

But by focusing on what is known about Iran's efforts to build or buy the technologies needed for a bomb, the report suggests that the country is working towards a relatively sophisticated device that could fit on board a medium-range ballistic missile — making it much more difficult to intercept and destroy than one delivered by an aeroplane.

The IAEA, an organization based in Vienna that monitors the nuclear facilities of NPT signatories, has been monitoring Iran's uranium-enrichment facilities for more than a decade. Those facilities use centrifuges to concentrate the fissile uranium-235 isotope, which makes up less than 1% of natural uranium. When enriched to 3.5–5% of the isotope, uranium can serve as a nuclear fuel; above 90%, it can be used to make a nuclear bomb.

Iran has insisted that its centrifuges are only making fuel-grade uranium, and it has given IAEA inspectors limited access to its facilities. But many questions remain about the nation's activities and plans to expand its enrichment programme, and most of the IAEA's reports on Iran have focused on those efforts.

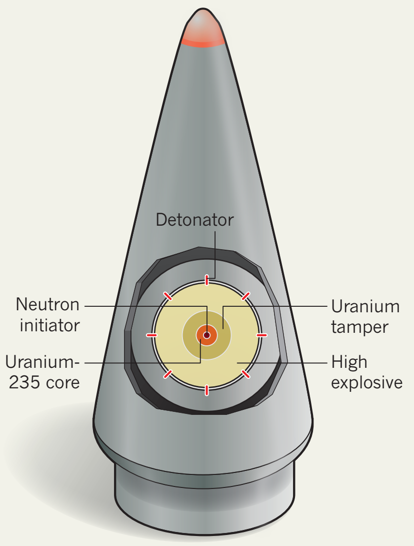
Political allegiances will always trump technical details.

Last week's report, however, described what is known about Iran's development of the technologies needed for a nuclear weapon. This includes techniques for shaping uranium metal, something that is not usually needed in civilian nuclear reactors but is required to make the precisely machined components that can power a nuclear explosion.

Iranian researchers have also been testing high explosives on tungsten, a dense metal that can serve as a surrogate for uranium. Such studies would be needed if they wanted to compress uranium to the critical mass needed for a self-sustaining nuclear reaction. And they have been looking at devices that, when compressed rapidly, produce bursts of neutrons that could trigger a nuclear chain reaction. The mass and shape of the materials tested seem to be designed to fit atop the Shahab-3, a medium-range ballistic missile developed by Iran.

IRAN'S NUKE

A report from the International Atomic Energy Agency suggests that Iran has been developing a nuclear weapon, based on a uranium implosion device, that would fit atop the Shahab-3 missile.



The agency says that much of the work has been done at the Malek-Ashtar University of Technology in Tehran, but Shahid Beheshti University and Amirkabir University of Technology have also been implicated. A year ago, Majid Shahriari, a nuclear physicist at Shahid Beheshti University, was killed in a bombing by unknown assassins (see *Nature* **468,** 607; 2010).

WEAPON DESIGN

The work outlined in the report suggests that Iran aims to create a weapon with an "implosion" design, says James Acton, a physicist with the Carnegie Endowment for International Peace, a non-profit think tank in Washington DC. The bomb would be detonated by high explosives surrounding a hollow sphere of highly enriched uranium (see 'Iran's nuke'). The core of the sphere would carry a small neutron initiator, possibly made of uranium and the heavy hydrogen isotope, deuterium.

When the explosives detonate in unison, they compress the sphere, squeezing the uranium to its critical mass. Near the point of maximum compression, the deuterium nuclei in the centre would fuse, releasing a burst of neutrons that would trigger the nuclear explosion. The device may also have an outer shell, or 'tamper', of low-enriched uranium, designed to hold the weapon together for a fraction of a second longer, further boosting its yield. Acton guesses that this kind of weapon could have a yield of around 10–30 kilotonnes of TNT equivalent, roughly the same as the bomb that fell on Nagasaki in Japan in 1945.

The advantage of the design is that it makes efficient use of uranium-235, so the device would be small enough to fit on a missile. The approach is similar to Pakistan's early warheads, which were also thought to be uranium-based. Unlike the warheads of the major nuclear countries, the Iranian design would not contain a second fusion stage, which can boost a weapon's yield into the 100-kilotonne range.

But even a simple implosion device is not an easy option. If the high explosives aren't detonated simultaneously, then the bomb will fail to explode properly and won't deliver its maximum yield. Many observers believe that this was the fate of North Korea's first nuclear test, conducted in 2006 (see *Nature* **443,** 610–611; 2006), although a second test in 2009 seems to have been more successful.

Iran's choice of uranium could also complicate its nuclear efforts. Most nuclear weapons use plutonium-239, because it captures neutrons better and emits more neutrons as it splits, giving it greater explosive power. The advantage of uranium over plutonium, however, is that it requires smaller production facilities, which are easier to hide.

Some analysts, including Vaez, are unimpressed by the IAEA's latest report, saying that much of the content dates from the turn of the millennium, and that it does not indicate how far the programme has progressed. But Vaez notes that "it is unprecedented in the scale and scope of the detailed information that it has bared to the public". That suggests to him that it may be a political push to encourage Russia and China to impose sanctions against Iran.

Acton agrees, although he adds that the report is unlikely to have that effect. For the countries that support Iran's right to enrich uranium for civilian use, political allegiances will always trump technical details, he says. "More evidence is not going to necessarily lead them to change their positions." ■

Life on the farm

Five years in, has a lofty experiment in interdisciplinary research paid off?

BY M. MITCHELL WALDROP

Gerald Rubin points to three jumbo coffee urns that stand near a dining area in the Janelia Farm main laboratory building, an elegant ribbon of glass and concrete overlooking the Potomac River valley near Ashburn, Virginia. "We figured it was worth spending $20,000 a year to provide high-quality coffee for free," Rubin explains, "because that way, people won't be tempted to make it in their labs."

It is true that the coffee is good. But then, everything about this US$300-million facility testifies to the deep pockets of the Howard Hughes Medical Institute (HHMI), the not-for-profit research-funding organization in nearby Chevy Chase, Maryland, that opened the Janelia Farm Research Campus five years ago as its first-ever intramural research facility.

More significantly, the campus embodies what Rubin calls "an experiment in scientific culture"[1]. Put world-class researchers in an environment that makes it easy to interact across disciplines — right down to chance encounters around the coffee urns (see 'Cultivating collaboration'). Then put them to work on a handful of grand scientific challenges — long-term, high-risk, high-payoff research that addresses some of the biggest questions in neuroscience. And use the HHMI's ample chequebook to free them from the distractions of conventional academic life. No administrative work, no teaching duties, no chasing tenure, no writing of grants. "This really is the ivory tower," says Rubin, who was named director of the campus in 2003 and has been involved with the facility since its inception.

The hypothesis is that this $100-million-a-year experiment will produce uncommonly great science. To Rubin, this means being much more than simply excellent. "I'd consider us a failure if, in 20 years, we come back and say, 'We recreated the Salk Institute'," he says, hastening to add that he considers the Salk, located in La Jolla, California, to be one of the best free-standing research facilities in the world. "The point is that we didn't need to build Janelia Farm to do that," he says.

A geneticist through and through, he has proposed that Janelia must eventually be able to pass the "deletion test": just as knocking out a gene can reveal its function, removing Janelia from the future scientific landscape should reveal the vital importance of its contributions to biology.

EARLY DAYS

Five years into Janelia Farm's life, however, Rubin admits that he has no hard evidence that it will ever pass that exceedingly ambitious test. Given the campus's long-term research focus — and the difficulties of creating a brand-new institute from scratch on former farmland — Rubin says that he does not expect the facility to start producing its best discoveries for another five or even ten years. So far, the experiment has proved only that the promise of well-financed, unfettered academic freedom can indeed entice high-quality researchers to move to a new facility.

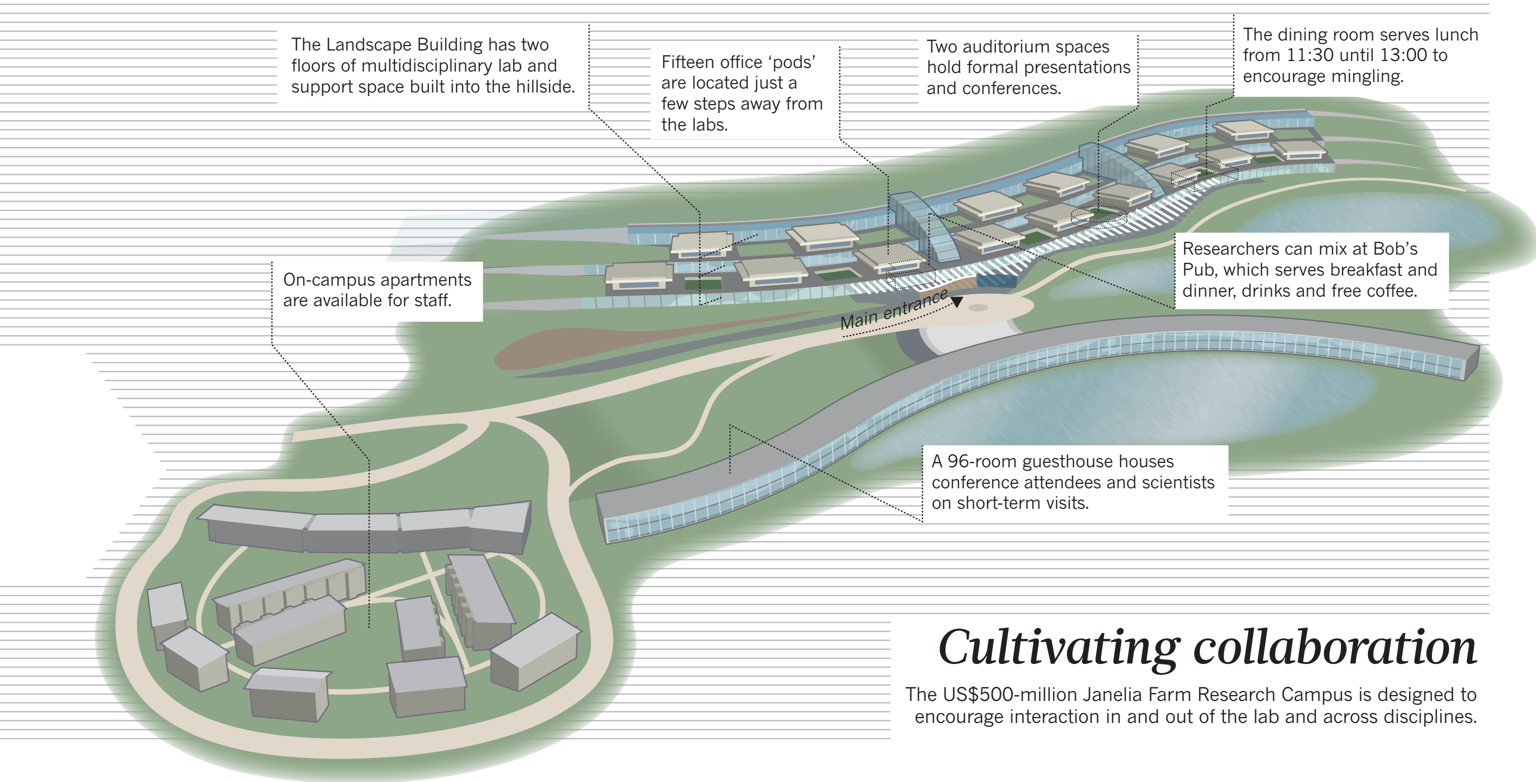
Still, the researchers' presence is reflected in the steady increase in publications from Janelia Farm labs (see 'Publications on the rise'), and in the growing respect Janelia is commanding from investigators who were initially sceptical of its lofty ambitions.

Getting even this far has been a major accomplishment, says Eric Kandel, a neuroscientist at Columbia University in New York who was one such sceptic. "No one has hit a home run yet," he says. "But the team is in place."

The idea behind Janelia Farm originated in 1999, when Thomas Cech, a biochemist at the University of Colorado in Boulder, had just agreed to become president of the HHMI and was looking to try something new. The HHMI was already funding hundreds of investigators at universities around the world. And with its endowment booming, Cech thought there must be some way for the organization to have a bigger impact on science than just funding a few hundred more. To help him work out how to do this, he recruited Rubin, then a geneticist at the University of California, Berkeley, as HHMI vice-president for biomedical research.

Rubin began by looking back at some of the wonderful experiences in his own career: summers as an undergraduate at the Cold Spring Harbor Laboratory in New York; a PhD at the Medical Research Council Laboratory of Molecular Biology (LMB) at the University of Cambridge, UK; and three years at the Carnegie Institution for Science Department of Embryology in Baltimore, Maryland. He wondered: what had made these places feel so great?

Rubin posed the question to many people, including veterans of non-biological institutes — notably Bell Labs in Murray Hill, New Jersey, which had been an innovation powerhouse before the 1984 break-up of its parent company, AT&T. The answers from all these places were surprisingly consistent, he says. Research groups were small, which promoted communication and mentoring. Group leaders were active bench scientists,



Cultivating collaboration

The US$500-million Janelia Farm Research Campus is designed to encourage interaction in and out of the lab and across disciplines.

not administrators or fund-raisers. Research was funded from within, so there was no need to chase grants. And no one got tenure, so that researchers could rotate through and ideas could stay fresh.

Many institutes had implemented some of these principles. Bell Labs and the LMB, despite widely divergent remits in applied physics and molecular biology, had implemented all of them in their glory days. But no one was doing so at the time, says Rubin — especially not the internal-funding part. And that, he argued, was the HHMI's great opportunity.

Cech and the HHMI trustees were sold on the idea from the beginning. But others were not. "I didn't see what was so special about recreating the LMB," says Kandel, recalling his early scepticism. "It wasn't clear to me what problems would be solved there, or even what fields they would be working in," he says. Many university investigators funded by the HHMI worried that this new initiative would end up cutting into their money. And most of them doubted that the HHMI would be able to persuade top-quality people to forgo tenure in established research centres and move to a farm outside Ashburn, a dormitory suburb an hour's drive from Washington DC.

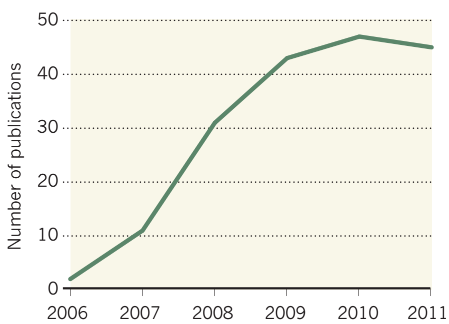
Also controversial was the initiative's proposed research strategy of focusing on a handful of grand challenges instead of tackling a wide range of biomedical problems. In 2004, the HHMI held a series of five workshops to determine what those grand challenges would be. One topic they settled on fairly quickly was technologies for biological imaging. "It was a great problem for us," says Rubin. Not only would it bring together physics, chemistry, biology and many other disciplines, he says, "it was going to be an enabling technology for so many areas, the way [DNA] sequencing had been".

A second major topic was understanding neural circuits and how they give rise to behaviour. This promised to fill a tremendous gap in neuroscience, says Kandel. "It was clear that we now understood neurons very well," he explains. "And we had imaging techniques to see how large areas of the brain are interconnected. But there was nothing to link the two." A new generation of techniques was poised to aid in this quest — most notably optogenetics, in which the activity of specific neurons can be tracked and manipulated using light, allowing researchers to work out the function of those neurons and how they connect.

But at the time, the application of optogenetics to neuroscience was still in its infancy. And there was always the chance that Janelia Farm researchers would spend 20 years deciphering the neural circuitry of, say, the fruitfly *Drosophila melanogaster*, only to discover that it had nothing to do with the human brain. Most people thought that the underlying principles and logic of the circuits would have been conserved throughout evolution, says Kandel, but still, "it was a very ballsy move".

PUBLICATIONS ON THE RISE

The number of publications from Janelia scientists has increased since 2006, with papers appearing in top-tier journals such as *Nature*, *Cell* and *Science*.



SOURCE: G. RUBIN

FERTILE SOIL

In October 2006, the HHMI officially opened Janelia Farm's main laboratory facility, dubbed the Landscape building for the way it winds for some 300 metres along the vast, open 'S'-shaped curve of a hillside. It's not exactly a warm and cosy place. The hallways on the upper two floors, where the inner glass walls open onto row upon row of laboratory benches, are so big and full of light that they feel a bit like airport concourses. They also seem to be constantly vanishing around the next bend, which can produce the disconcerting sense that one is stepping off into infinity.

But none of that bothered Julie Simpson when she arrived in the summer of 2006, fresh from a postdoctoral appointment in neuroscience at the University of Wisconsin–Madison. She had been eager to get to Janelia Farm ever since hearing Cech give a talk about its philosophy — so eager, she says, that when she moved in, "I had to wear a hard hat because they didn't even have finished floors in my lab yet!"

And she hasn't been disappointed. "I really like being part of a community focused on a particular problem," she says. "Every talk is relevant, and every colleague is interested in the same basic thing" — albeit with different perspectives and lots of productive arguments. Simpson leads a group using optogenetics to

M. STALEY/JANELIA FARM

The main building of the Janelia Farm Research Campus in Ashburn, Virginia, houses laboratory space within expansive windows and interior glass walls.

trace the neural circuitry that gives rise to specific, hard-wired responses — grooming behaviour, for example — in *D. melanogaster*. Comparing the detailed structure of many such circuits, she says, should begin to reveal the general principles behind them[2].

Another early recruit was Eric Betzig, a physicist who had begun working at Bell Labs during the late 1980s, when the facility was still investigating everything from neuroscience to antimatter. Impatient with the steadily declining role of basic research at the labs, Betzig left to work at his father's machine-tool company in 1994. Soon after deciding to get back into research, he joined Janelia Farm in 2006 to work on a broad range of imaging technologies.

"What attracted me was the same thing that had attracted me about Bell," he says. "All the resources you need, and no pressure to publish. I won't do science if I can't do it under those terms." Before arriving at Janelia, Betzig had been developing an imaging technique, called photoactivated localization microscopy, or PALM, that he created in collaboration with Harald Hess, a physicist who is now also at Janelia Farm. It uses image-processing algorithms pioneered by astronomers to detect the position of single fluorescent molecules with nanometre accuracy[3]. Since starting at Janelia, Betzig has developed three more imaging techniques designed to help biologists peer deeper than before into the layers of living cells, with higher resolution and with less damage to them in the process.

Janelia Farm is definitely not for everybody, says Rubin. "I've tried to make it irresistible for a small fraction of people," he says — researchers who are confident enough to go without tenure, who don't mind the remote location or the six-person limit on group size, and who do want to work with their own hands.

Plenty of researchers do seem to fit that description. At present, Janelia Farm has 20 research-group leaders, who are evaluated for renewal on a five-year cycle. The first round of evaluations begins next spring — a process Simpson calls "terrifying", if only because it's new and no one knows for sure how it will work. In addition, 26 fellows at various stages of non-renewable five-year stints are working at the facility, as are more than 100 visiting scientists. With additional group members and support staff, the total number of employees at Janelia Farm comes to 424.

Yet the place can still feel almost empty. "That's the first thing you notice," says Robert Tjian, a biochemist who was once Rubin's colleague at Berkeley and who succeeded Cech as HHMI president in 2009. "We probably have another 100 to 150 scientists to put in the building before it's even close to being full."

KEEPING CREATIVE

Those spaces are empty in part because it has taken longer to recruit scientists than Rubin initially thought. The most common problem for prospective recruits is the six-member limit for research groups, he says. The remote location has been a smaller hurdle than feared, but it has been a factor.

Rubin and his colleagues have done their best to fight that isolation. Janelia's visitor programme brings scientists in for weeks or even months at a time, and the campus hosts about a dozen scientific conferences every year. Nonetheless, the remoteness of the campus is a major ongoing challenge, says Carla Shatz, a neuroscientist who directs the interdisciplinary Bio-X programme at Stanford University in California, and who serves on the Janelia Farm advisory committee. "Janelia has to think about how it can create excitement and creativity and innovation in a location that doesn't have ready access to a medical school, an engineering school, biology, physics and chemistry departments or an undergraduate student body," she says.

A related challenge is that of keeping the research programme intellectually fresh. The tight focus on imaging and neurocircuitry has been useful in the initial phase, says Tjian, "so that everyone understood what Janelia Farm was". But in the long run, he says, the place will stagnate unless it can broaden out and give scientists the freedom to follow new opportunities as they arise.

That is another reason Janelia still has so many empty spaces: they allow for expansion. On the basis of a series of planning workshops held earlier this year, Rubin has begun to hire group leaders in the fields of cell biology, evolutionary biology, structural biology and chemistry — each of which has some overlap with neurocircuitry but will also extend the campus's programme in new directions.

Looking back over Janelia Farm's first five years, Rubin says he is very satisfied. "We showed that we could go from an empty building to a functioning lab, that we could hire first-rate people, and that they could come here and work on interesting problems," he says. "All the things people said we were certain to fail at, we accomplished."

But what happens next? Can Janelia Farm do 'great science' during the next 5 to 10 years? Will it pass Rubin's deletion test? Can it rewrite the introductory biology texts (Cech's favourite definition of great science), or foster "a couple of programmes that create a whole new direction" (Tjian's favourite)?

That is the great unanswerable question. As Simpson says, "you can't engineer great science. You just have to create the conditions that make it possible, and see what happens." ■ SEE CAREERS P. 433

M. Mitchell Waldrop *is a features editor for* Nature *in Washington DC.*

1. Rubin, G. M. *Cell* **125,** 209–212 (2006).
2. Simpson, J. H. *Adv. Genet.* **65,** 79–143 (2009).
3. Betzig, E. *et al. Science* **313,** 1642–1645 (2006).
4. Pfeiffer, B. D. *et al. Genetics* **186,** 735–755 (2010).

LOST WORLD

Did a giant impact 200 million years ago trigger a mass extinction and pave the way for the dinosaurs?

BY ROFF SMITH

It takes a little fiddling, a few missed turns on the old, labyrinthine lanes and a good deal of folding and unfolding of an Ordnance Survey map bought that morning, but eventually Paul Olsen and Dennis Kent manage to locate the unmarked access path that leads through the woods to a desolate stretch of shoreline on Lavernock Point, a wild, cliff-lined promontory south of Cardiff in Wales, UK.

Olsen and Kent park their rental car in a muddy lay-by. A slow rain begins to fall and low peals of thunder can be heard in the distance. Grumbling good-naturedly about the British climate, the two geoscientists shrug into their parkas, sling their rucksacks over their shoulders and start down the slick path for another afternoon of getting cold, wet and muddy in the search for clues to mysterious events that wiped out much of life on Earth 200 million years ago and allowed the dinosaurs to take over the world.

How the dinosaurs' mighty reign ended has been fairly well established: a catastrophic asteroid impact near Chicxulub on the Yucatan Peninsula in Mexico, just over 65 million years ago, is widely credited with bringing the age of the dinosaurs to a close and ushering in the age of mammals.

Olsen and Kent, who both work at Columbia University's Lamont-Doherty Earth Observatory in Palisades, New York, have long speculated that with Chicxulub, history might have been repeating itself: that another asteroid, 135 million years earlier, could have wiped out, or at least had a hand in wiping out, much of the late-Triassic flora and fauna. This would have allowed dinosaurs to spread around the globe during the subsequent Jurassic period (200 million to 145 million years ago), evolve enormous bodies and dominate the planet until the next great impact catastrophe.

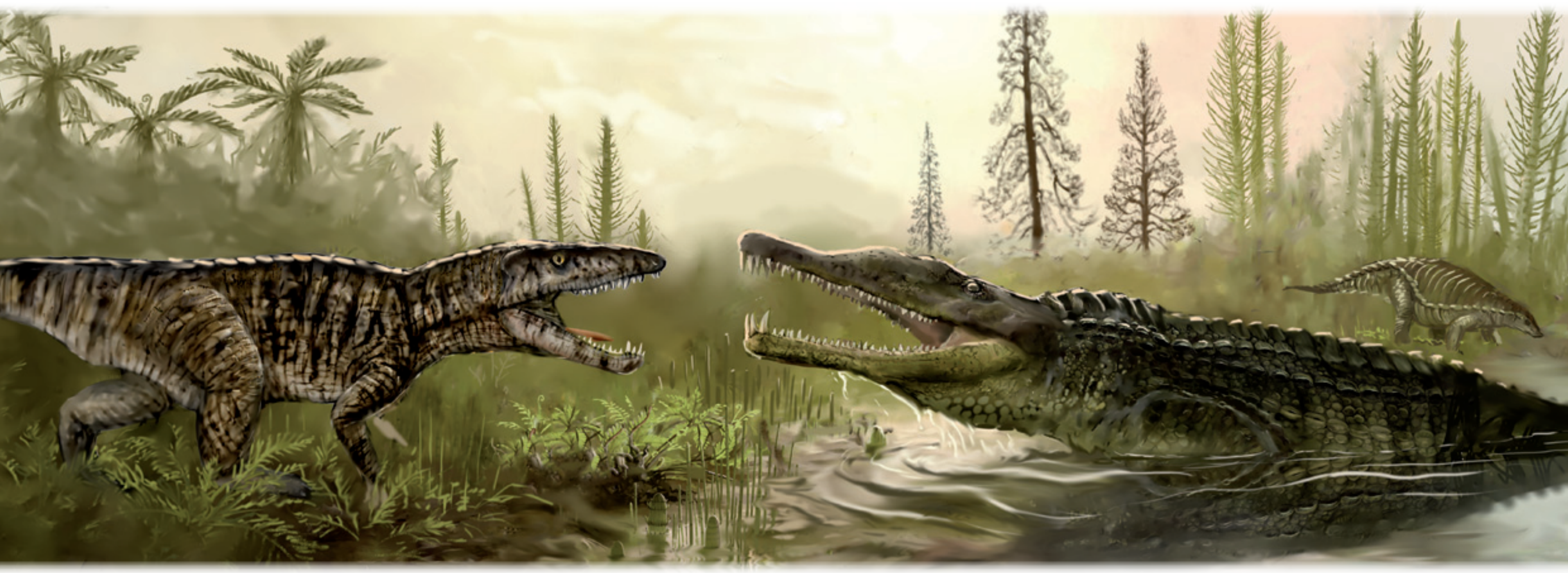
SUDDEN DEATH

It is certain that something drastic did happen at the end of the Triassic, because in the space of a few thousand years, half of all genera known to have existed at the time suddenly vanish from the fossil record. At sea, 20% of all families abruptly disappear, including an entire class of creatures — the eel-shaped conodonts. On land, the death toll was even higher. It was one of the greatest mass extinctions in Earth's history and, at this vast distance in time, one of the least understood. "The only thing anyone can say with any certainty about the Triassic–Jurassic mass extinction is that it happened," says Olsen. "Whatever it was that caused it, it happened so swiftly that most life forms never had time to adapt and evolve to meet the changes."

The prevailing view among scientists these days is that the extinctions were caused by massive volcanic activity associated with the break-up of the super-continent Pangaea. The series of eruptions created a vast geologic formation called the Central Atlantic magmatic province (CAMP; see 'End of an era'). "We are talking here about volcanic activity on a scale many thousands of times greater than anything ever witnessed by humans," says Gregory McHone, an independent Canadian geologist who has spent much of his career investigating the CAMP volcanic event and building a convincing case for its involvement in the Triassic–Jurassic mass extinction.

The flood-basalt eruption of the Icelandic volcano Laki in 1783 provides researchers with a scaled-down model of just how bad things might have been in the late Triassic. An outpouring of sulphurous gases from Laki created

VICTOR O. LESHYK



Extinctions at the end of the Triassic period killed off top predators such as *Redondavenator* (left), seen in an artist's impression confronting an aquatic phytosaur.

haze that cooled the planet and caused widespread crop failures and famine. It ultimately contributed to the deaths of an estimated 6 million people, says McHone. But disastrous as Laki's eruption was, it belched up just 15 cubic kilometres of basalt. The CAMP events produced 2 million cubic kilometres or more, in a series of pulses that alternated between cooling the climate with sulphurous haze and warming it with massive emissions of carbon dioxide and methane. The oceans grew acidic and parts became starved of oxygen, while on land a surge in lightning sparked extensive fires[1].

"It happened so swiftly that most life forms never had time to adapt and evolve."

Many of Earth's life forms simply couldn't recover from that succession of body blows, says McHone.

It's a plausible theory, as Olsen and Kent both readily concede — even, perhaps, the most likely one. All the same, the CAMP theory leaves a lot of unanswered questions, not least of which is the suddenness of the extinctions. The late-Triassic eruptions span hundreds of thousands of years, but the die-offs seem swift in the fossil record. And what of the 'fern spike'? Late-Triassic sediments on the US east coast contain huge quantities of fossilized fern spores.

Ferns are usually the first plant to appear after a natural disaster, says Kent. "If you look in the fossil record you see a sudden massive spike in ferns just around the time of the extinctions, but demonstrably before the great basalt flows — at least in our part of the United States." In other spots, the extinctions seem to coincide with the oldest basalt layers, within the errors of the dating conducted so far.

"The only way we are ever going to unravel this mystery is to work out a timeline, as precise as we can make it, of all the various events around the world that led up to it," says Kent. Pursuit of that timeline has taken Olsen and Kent on a global quest, from North Carolina to Nova Scotia, Canada; China to Germany, Italy, Austria and the High Atlas Mountains in Morocco; and now to Wales.

JUST IN TIME

Olsen has been thinking about the possibility of an impact connection to the Triassic–Jurassic extinctions for more than 20 years[2,3], but one of the biggest drawbacks to the asteroid theory is that until recently nobody had found any evidence of such a catastrophe occurring around the time of the extinctions. Then, last year, French and German research teams re-dated a badly eroded structure left by a massive impact near Rochechouart, in western France[4]. Previous work had put the impact at around 214 million years ago, long before the extinction event. But the revised date of 199 million to 203 million years ago overlaps with the extinctions, which have been dated to 201.4 million years ago[5].

The authors of the paper also suggested that the enormous shock waves generated by the 2-kilometre asteroid, as it slammed to Earth at more than 25 kilometres per second, could account for unexplained ripples and disturbances in the late Triassic limestone and shale beds in western Britain — sedimentary beds that coincide with the mass-extinction event. "When I read that," says Olsen, "I decided it was time to come over here and take a much closer look."

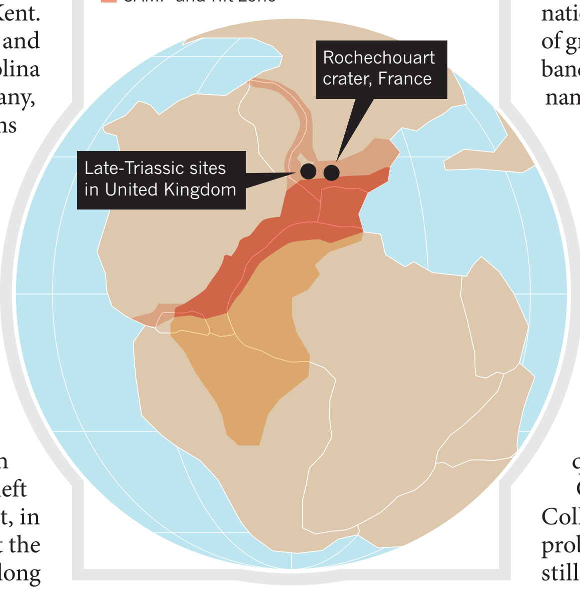
As Olsen scrambles along the base of the Welsh cliffs, his stories bring to life the Triassic world. There were the monkey lizards with their beaky faces, long arms and grasping tails; crocodilian creatures that trotted along like dogs; and shallow tropical seas teeming with sharks, right where this shingle beach lies today. "Then suddenly it all came crashing to an end," says Olsen.

Olsen's fascination with this lost world goes back to 1968, when he was a 14-year-old in Livingston, New Jersey, and heard that dinosaur footprints had been found in the rocks of the nearby Roseland quarry. He and a school friend hopped on their bikes, pedalled over to the quarry and found that there were fossils everywhere.

End of an era

The close of the Triassic period, 200 million years ago, was a tumultous time. The supercontinent Pangaea was starting to split apart, massive volcanic eruptions occurred, a large asteroid hit the planet and there was a mass extinction.

- Rift zone
- Central Atlantic magmatic province (CAMP) zone
- CAMP and rift zone



By the time the two friends were in their second year of high school, they had catalogued thousands of fossils and tracks of reptiles from the late Triassic and early Jurassic. They became so involved that when the quarry and its treasures were to be sold off and developed into housing units, the teens mounted a publicity campaign to save it.

Soon, *Life* magazine was on the phone and their campaign had captured national attention. Olsen even made a cast of a footprint left by a fearsome three-toed beast, and sent the fibreglass model to then-US president Richard Nixon.

The publicity garnered by this sheer chutzpah helped to protect the fossil-rich site, and the fibreglass footprint went on to find a place in Richard Nixon's presidential library.

Olsen's trip through western Britain with Kent retains some of that sense of schoolboy enterprise and science on the human scale: the two researchers bicker amiably over which way to go at the crossroads; stop off in the local supermarket to pick up more plastic sandwich bags for their rock specimens; use a self-modified, battery-powered drill to take core samples; and even get scolded by the waspish landlady at their guest house when they clomp in with muddy boots at the end of the day.

EXPLOSIVE FORCE

At Lavernock Point, the scientists ignore the rain and thunder as they set about their work. Olsen makes his way to the base of the cliffs and points to a layer of buff-coloured limestone at about waist height. "Right there is where it happened: there's the extinction line," he says. To a layman, it is innocuous: just another of the alternating bands of limestone and shale that are brightened here and there by clumps of flowering purple valerian. But close examination reveals sand-filled cracks and deposits of grainy, irregularly sized material — disturbances that might have been the work of a tsunami or an extraordinary earthquake.

The disturbed layers here and at other late-Triassic sites in the United Kingdom have a distinctive orientation, all angled as if the source of the unrest was in the vicinity of Rochechouart, says Olsen.

It certainly would have been an awful day in this part of the world when the Rochechouart crater formed. The researchers who re-dated it estimate[4] that the impact would have generated an earthquake up to magnitude 11.5 — 100 times more powerful than any quake in recorded history.

Gareth Collins, a geoscientist at Imperial College London, says that the quake would probably have been much smaller, although still massive. Collins and researchers from

AP/JIM ROSS

Paul Olsen inspects rocks along a beach in Wales to trace what caused one of the biggest mass extinctions on Earth.

Purdue University in West Lafayette, Indiana, have developed an online calculator to model the effects of impacts. Their algorithm paints a vivid picture of what would have been going on at this spot in Wales after the asteroid hit with the explosive force of more than one million megatonnes of trinitrotoluene (TNT). Even at a distance of more than 600 kilometres, the beach would have endured hurricane-force winds and a hail of debris.

That debris would have carried a chemical signature of the impact that might still reside in the layers of sedimentary rocks. Stratigrapher Stephen Hesselbo and geochemist Ken Amor, both at the University of Oxford, UK, have joined Olsen and Kent on their outing, and are taking samples to analyse with a mass spectrometer. They will measure the ratios of chromium isotopes, and look for one that is characteristic of meteorites. Olsen will also send samples from this location to another lab, to be searched for other tell-tale impact markers.

If those tests detect an extraterrestrial signal in the late-Triassic sedimentary layers in Wales and elsewhere, it will be the first substantive link between an impact and the Triassic–Jurassic mass-extinction event. But a coincidence in time won't prove that an impact was the cause. "The two things might occur at approximately the same moment but have nothing whatever to do with each other," says Olsen.

More troublingly, the Rochechouart impact doesn't seem to have been anywhere near big enough to have accounted for the mass extinctions around the globe — at least, not on its own. The 25-kilometre-wide buried crater may have been 40 to 50 kilometres across originally, but it is just a pockmark in comparison with the 180-kilometre-wide scar at Chicxulub. "Based on our estimates, Rochechouart is quite small in terms of global environmental consequences," says Collins.

PICKING UP THE PIECES

For now, says Olsen, it is too early to make any definitive statements about what Rochechouart did or didn't do. That will take much more data from late-Triassic sites around the world. This year alone, he has crossed the Atlantic three times to collect samples from the United Kingdom and Morocco. He is back in the Atlas Mountains this week, to examine yet more sites. He and his colleagues are not only looking for signs of an extraterrestrial impact in the sediments, but are also searching for other clues, such as the extinction layer and a chemical signature that could be linked to the CAMP eruptions. All those data, says Olsen, will help the team to sort out the relative timing of events around the world, and to create a fuller picture of what happened and how life responded.

He suspects that Rochechouart may have been "a piece of a much bigger puzzle". Perhaps it was one of a series of asteroids that hit around the same time. Alternatively, a lone French crash might have been the final straw for a world already reeling from volcanic eruptions. Or the impact may have come first, weakening ecosystems enough that when the eruptions started, life took a nosedive.

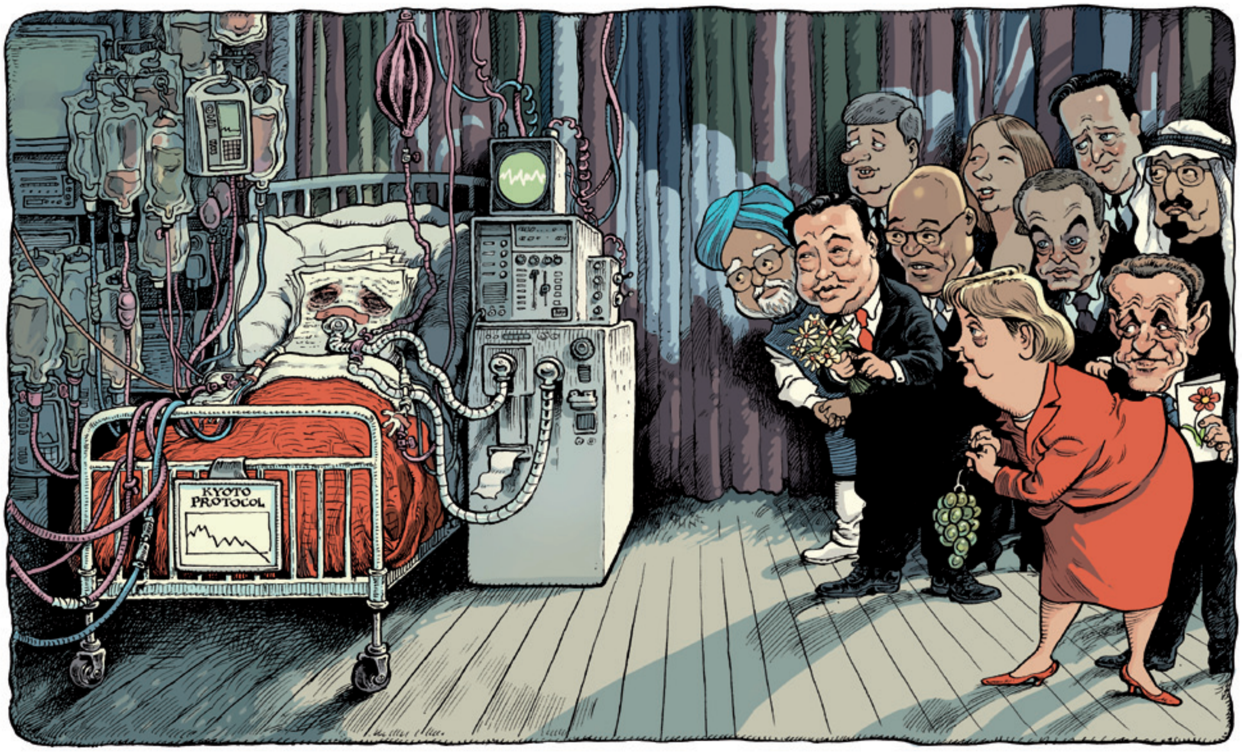
Teasing out the answer will take some time, says Olsen, as he trudges, wet and muddy, back up the trail at the end of another wearying afternoon in the cold Welsh rain. "There is no easy way to do this," he says. "But I believe that eventually we will be able to put the pieces together and know what happened and why." ■

Roff Smith *is a freelance writer based in Hastings, UK.*

1. Belcher, C. M. *et al. Nature Geosci.* **3,** 426–429 (2010).
2. Olsen, P. E., Shubin, N. H. & Anders, M. H. *Science* **237,** 1025–1029 (1987).
3. Olsen, P. E. *et al. Science* **296,** 1305–1307 (2002).
4. Schmieder, M., Buchner, E., Schwarz, W. H., Trieloff, M. & Lambert, P. *Meteorit. Planet. Sci.* **45,** 1225–1242 (2010).
5. Schoene, B., Guex, J., Bartolini, A., Schaltegger, U. & Blackburn, T. J. *Geology* **38,** 387–390 (2010).

COMMENT

DAVID PARKINS



Letting go of Kyoto

A preoccupation with binding commitments blocks progress in climate-change negotiations. It is time to correct course, says **Elliot Diringer**.

When government representatives gather for another round of United Nations climate-change negotiations later this month in Durban, South Africa, they face a familiar thicket of issues. Yet for many — and, no doubt, for headline writers around the world — one stands above all the rest: the survival or death of the Kyoto Protocol. Kyoto's emission targets expire at the end of 2012, making Durban the last chance to set new targets in time to avoid a 'commitment gap'.

Kyoto will probably emerge from Durban alive, but just barely. This should not be cause for alarm. Although the protocol remains an important emblem of multilateralism, it has become, in reality, more of an impediment than a means to genuine progress. More important than ensuring Kyoto's long-term survival is building something better to take its place.

Durban affords an overdue opportunity to honestly reconsider what we can expect the UN climate process to deliver, and when. With the start of the Kyoto negotiations 16 years ago, the international community decided that legally binding commitments were the answer to climate change. A binding-or-nothing mentality has held sway ever since, and the result often has been 'nothing'.

Although it has been obvious for some time that most of the developed world is unwilling to one-sidedly assume new binding targets, many developing countries will arrive in Durban insisting on precisely that. Without a compromise, the outcome may be less than nothing. It might, in the worst case, be the unravelling of the entire enterprise.

The more sensible course is an incremental one. Modest successes were achieved at last year's climate-change negotiations in Cancún, Mexico. The parties should build on that with further steps to strengthen the regime; they should also declare their intent to work towards binding commitments, while acknowledging that this will take time. Meanwhile, governments and climate advocates must work at home to build domestic support for strong national action. Without that, future international commitments will mean little, whether binding or not. ▶

➲ **NATURE.COM**
For more ideas on what might come after Kyoto, see:
go.nature.com/ru5wk3

In Durban, governments will again be challenged by the same two fundamental issues that dominated the start of the global climate effort two decades ago. One is governance. Is the best approach a binding top-down treaty with sanctions for non-compliance, a loose bottom-up arrangement with countries free to define their own voluntary commitments, or something in-between? The second is fairness. How is effort against this quintessentially global challenge equitably apportioned among countries whose degrees of responsibility and capacity vary so widely, and are continually evolving?

FIRST PRINCIPLES

The 1992 UN Framework Convention on Climate Change took a first stab at both. On fairness, it established the principle that countries should act "in accordance with their common but differentiated responsibilities and respective capabilities". Applying that principle, it set specific obligations for developed countries only — returning their greenhouse-gas emissions to 1990 levels by 2000. But this was simply an "aim", not a binding target. As to the ultimate shape of the regime, the convention left the door wide open.

It soon became evident that most developed countries would miss this goal, and in 1995 the parties launched a new round of talks that led to the 1997 Kyoto Protocol. They agreed right off that new commitments would apply to developed countries only. And, inspired in part by the success of the Montreal Protocol on ozone-depleting substances, they decided that this time the targets would be legally binding. (The prescribed consequences for non-compliance, however, are technically not binding — illustrating the many shades of grey associated with the 'binding' concept.)

It took until 2005 for Kyoto to win enough ratifications — notwithstanding its renunciation by the United States — to enter into force. In that time and in the years since, the emissions picture has shifted dramatically. Global greenhouse-gas emissions are up 25% since 1997. China has overtaken the United States as the world's largest annual emitter. Collectively, emissions from developing countries are now 58% of the total, and rising fast.

Against this backdrop, it is no surprise that countries such as Japan, Canada and Russia adamantly refuse to assume new binding targets unless the other major economies at present outside Kyoto's reach — most notably, the United States and China — do so as well. And for now, the odds of that happening are nil.

Yet for many, binding commitments remain a holy grail. This produced a near disaster two years ago at the Copenhagen meeting, where the widely held but wholly unrealistic expectation of a binding outcome was destined to go unmet. World leaders managed to hash out a political deal, the Copenhagen Accord, but in the final vitriolic hours, a handful of parties blocked its formal adoption. To both those in the room and those watching from afar, the UN climate process seemed to teeter. Another go-round like that in Durban could push it over the edge.

A key premise of the Kyoto experiment was that binding international commitments would drive national efforts. Yet outside Europe, where concern about climate change has always run strongest, there is little evidence that this is true. A prime counter example is Canada, where emissions are now 17–30% above 1990 levels (depending on whether land-use emissions are counted), despite a binding commitment to reduce them to 6% below.

Where ambitious national efforts have emerged, two other drivers seem more influential: political will and economic self-interest. Australia is arguably a case of the former. Heavily reliant, like Canada, on natural resources and energy-intensive exports, Australia's last government fell when it tried to push through emissions trading. But the new minority government — a coalition including the Green party — recommitted to the issues and just this month enacted an ambitious carbon-pricing scheme.

The mercantile motive, meanwhile, is nowhere more evident than in China, which has quickly dominated the emerging clean-energy market and now produces nearly 50% of the world's wind turbines and solar panels. China will also soon introduce emissions trading at the regional level.

In most cases, economic motive and political will both play a part. Germany and the United Kingdom are going beyond European Union (EU) mandates with 2020 emission targets of 40% and 34%, respectively, below 1990 levels. South Korea has devoted most of its 2009 US$38-billion economic stimulus to green growth, including energy efficiency and renewables. It and some other developing countries, including Brazil, India and South Africa, are fashioning or implementing market-based policies to drive efficiency or reduce emissions. Where neither political will nor competitive drive has yet taken hold, as in the United States, investment and action unfortunately lag.

> ***"Strong, durable agreements don't typically spring forth fully formed."***

If the principal drivers of action are domestic, do international commitments matter? Yes. In the long term, Kyoto's adherents are right: emissions commitments should be binding. Strong, sustained action to preserve a global good requires confidence that all are indeed contributing their fair share. But we need to be more realistic about how and when we get there.

Fortunately, if governments are prepared to look beyond Kyoto, they can find in last year's Cancún Agreements the seeds of a more viable successor. That pact gave the essential elements of the Copenhagen Accord the UN imprimatur, and offered countries the opportunity to pledge explicit targets or actions for 2020. More than 80, including all the major economies, have now done so.

This time, the numbers were set unilaterally, not negotiated as in Kyoto; pledges came from both developed and developing countries; and they are voluntary, not binding. In other words, even as the Kyoto negotiations have dragged on, a parallel 'bottom-up' framework has begun to take shape.

As yet, it is hardly adequate. To begin with, the 2020 pledges are too weak to put countries on track towards limiting global warming to 2 °C, the goal set in Copenhagen and affirmed in Cancún. Beyond that, the operational elements of Cancún — including a new Green Climate Fund for developing countries, stronger reporting and scrutiny of countries' actions, and new adaptation and technology mechanisms — are mere shells, with a raft of details still to be agreed on.

SMALL STEPS

In Durban, parties should indeed set their sights towards eventual binding commitments. But they should focus primarily on the more prosaic nuts and bolts of strengthening transparency and support for developing countries. However incremental, such steps will get us further than a recurring cycle of false expectation and failure.

For the Kyoto Protocol itself, the likely outcome is some sort of half-measure. A leading option is to set new emission targets through a 'political' second commitment period, which can be approved outright by ministers gathered in Durban, rather than a legally binding amendment to the protocol, which would have to be brought home and ratified, a long and difficult process for many governments. Even if joined by only the EU and a handful of others, such a life-support mechanism would avert a blow-up, and buy time to build a sounder alternative.

Looking across the multilateral landscape, it is clear that strong, durable agreements don't typically spring forth fully formed — they evolve over time. Kyoto was a bold attempt to short-circuit the process. The real tragedy is not its demise, but that the binding-or-nothing mindset has in the meantime kept us from pursuing other multilateral means of tackling climate change. Durban is a chance to correct course. ■

Elliot Diringer *is executive vice-president of the Center for Climate and Energy Solutions in Arlington, Virginia, USA, formerly the Pew Center on Global Climate Change.*
e-mail: diringere@c2es.org

Good governance for geoengineering

Phil Macnaghten and **Richard Owen** describe the first attempt to govern a climate-engineering research project.

Climate-engineering research must have strong governance if it is to proceed safely, openly and responsibly[1,2]. But what this means in practice is not clear. The Stratospheric Particle Injection for Climate Engineering (SPICE) study demonstrates the difficult judgements involved. As chairman of the panel that supported decisions by the UK Engineering and Physical Sciences Research Council (EPSRC) as to whether and how this project should proceed (P.M.), and the architect of the project's governance process (R.O.), we draw lessons from these challenges.

In mid-September 2011, SPICE announced the go-ahead for the United Kingdom's first field trial of climate-engineering technology. SPICE aims to assess whether the injection of sulphur particles into the stratosphere would mimic the cooling effects of volcanic eruptions and provide a possible means to mitigate global warming. An equipment test — spraying water at a height of 1 kilometre — was proposed (see 'SPICE field trial'). No climate engineering would result from the test, but response to the announcement was dramatic, and the project was soon at the centre of a storm of criticism.

CAREFUL REVIEW

On 26 September 2011, the EPSRC, one of the study's main funders, postponed the trial after a review. Later the same day, the council received a letter and open petition[3], also sent to UK energy and climate-change secretary Chris Huhne and signed by more than 50 non-governmental organizations (NGOs) and civil-society organizations, demanding that the project be cancelled. The signatories saw the research as a first, unacceptable step towards a fix that would deflect political and scientific action away from reducing greenhouse-gas emissions. Others, by contrast, saw the research as urgently needed to find possible ways of coping with climate change[4]. The question at the heart of this debate was: should work in this controversial field proceed at all, and if so, under what conditions?

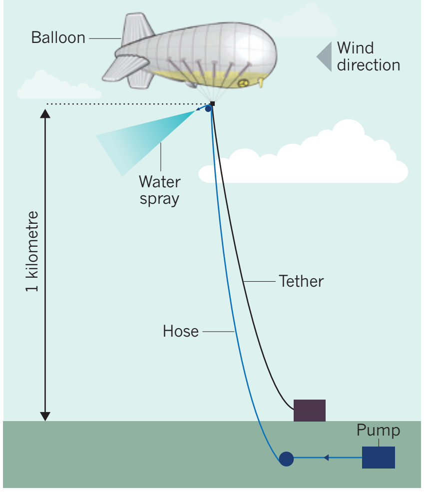
The strong feelings about the first test of SPICE's equipment show how important it is to have robust governance, and for scientists and funders to ensure that the public and other parties are consulted at the earliest opportunity. This is an unfamiliar and difficult process, but it is crucial for the evaluation of climate-engineering approaches.

SPICE was conceived in March 2010 at an EPSRC interdisciplinary workshop, at which researchers were invited to develop innovative geoengineering proposals. The project's funding incorporated field testing, but release of money was conditional upon it passing a 'stage-gate' review — a governance process in which funding for each phase of research and development is preceded by a decision point. To pass the review, SPICE scientists were required to reflect on the wider risks, uncertainties and impacts surrounding the test and the geoengineering technique to which it could lead — solar-radiation management.

On 15 June 2011, the stage-gate panel (including atmospheric scientists, engineers and social scientists, as well as an adviser to an environmental NGO) evaluated the SPICE team's response to five criteria for responsible innovation. These were that: the test-bed deployment was safe and principal risks had been identified, managed and deemed acceptable; the test-bed deployment was compliant with relevant regulations; the nature and purpose of SPICE would be clearly communicated to all relevant parties to inform and promote balanced discussion; future applications and impacts had been described, and mechanisms put in place to review these in the light of new information; and mechanisms had been identified to understand public and stakeholder views regarding the predicted applications and impacts.

SPICE FIELD TRIAL

Water sprayed through a 1-kilometre-high hose will test equipment with potential for climate engineering.



Recognizing the efforts of the SPICE team, the panel concluded that although the first two criteria had been met, more was required on the remaining three. It asked the team to develop a revised communications plan to inform further public debate, a review of the risks and uncertainties of solar-radiation management — including social, ethical, legal and political dimensions — and a thorough process of engagement with stakeholders.

The test bed was delayed by EPSRC in September to allow the team to undertake these outstanding actions. When the panel reconvenes, it will independently assess a revised response; until then, the project remains under review.

LESSONS LEARNED

Aspects of SPICE's governance could have been improved. The framework should have been in place before the project's conception; the test date should not have been announced until the stage-gate criteria had been met; and the structures and resources to support the social research should have been in place earlier. Even now, the decision on whether to proceed will not be easy. There are few right or wrong answers to the many questions about climate engineering. But it is vital that we make space to listen to and discuss these questions, and that the debate transparently influences the decisions that are taken.

For geoengineering technology to progress, its developers must be mindful of wider impacts from the outset; and it must proceed under robust governance mechanisms. The SPICE responsible-innovation framework is one evolving approach to achieving it. ■

Phil Macnaghten *is professor of geography at Durham University, UK.* **Richard Owen** *is chair in responsible innovation at the University of Exeter Business School, UK.*
e-mail: p.m.macnaghten@durham.ac.uk
R.J.Owen@exeter.ac.uk

1. Royal Society working group *Geoengineering the Climate: Science, Governance and Uncertainty* (Royal Society, 2009) available at http://go.nature.com/zxpwun
2. Rayner, S., Redgwell, C., Savulescu, J., Pidgeon, N. & Kruger, T. *Memorandum on Draft Principles for the Conduct of Geoengineering Research* (House of Commons Science and Technology Committee Enquiry into The Regulation of Geoengineering; 2009).
3. http://www.handsoffmotherearth.org/hose-experiment/spice-opposition-letter/
4. Nurse, P. Letter to *The Guardian* 8 September 2011 available at http://go.nature.com/efnybg

Neanderthals were wary xenophobes who had considerable empathy and liked slapstick humour.

RECONSTRUCTION: KENNIS & KENNIS/PHOTO: J. MCNALLY

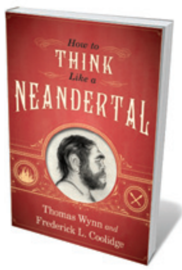
NEUROSCIENCE

Neanderthals in mind

Clive Gamble relishes the inside story on the cognitive abilities of our fossil relatives.

Wondering what went on in the heads of Neanderthals has rarely produced positive thoughts. H. G. Wells set the bar low in his short story *The Grisly Folk* in 1921, writing: "We cannot conceive in our different minds the strange ideas that chased one another through those queerly shaped brains." Wells's hatchet job was effective. Other authors have offered sympathetic alternatives, such as Isaac Asimov's 1958 short story *The Ugly Little Boy*. But the idea of a 'thinking Neanderthal' has become an evolutionary oxymoron on a par with 'military intelligence' and 'airline food'.

Yet cognition certainly took place in the Neanderthal brain — the largest in human evolution, housed in a long, distinctively shaped skull. In *How to Think Like a Neandertal*, archaeologist Thomas Wynn and psychologist Frederick Coolidge provide one of the most rounded portraits yet of a fossil human. The book covers familiar areas — diet, symbolism and language — but also includes innovative assessments of Neanderthals' capacity to tell jokes, and even speculations on what they might have dreamed about. The authors use the Neanderthals as a means of discussing the evolutionary reasons for such cognitive abilities as humour and deception.



How To Think Like a Neandertal
THOMAS WYNN AND FREDERICK L. COOLIDGE
Oxford University Press: 2011. 224 pp. £16.99, $24.99

We have learned much about Neanderthals in the past 150 years. They were powerfully built and top carnivores. Their stone tools are found across Eurasia. We know from their genome sequence that the last common ancestor of Neanderthals and ourselves lived some half a million years ago. They became extinct in southern Spain as recently as 30,000 years ago.

Yet understanding these remarkable people is hard. There is a veneer of common prejudice about primitives and progress that first has to be stripped away. Then there is the awkward fact that what they made and left behind is unimpressive. Their tools changed little over time and space. They had fire and, on occasion, a way of disposing of their dead that accords with what we understand as burial. But they made no art in the form of painting or carving, just a few perforated and pigmented shells.

Archaeologists look for cultural diversity and innovation when they judge people in the past. Sadly for the Neanderthals, they rubbed shoulders in Europe with modern humans — our direct ancestors — who had the latest technology, charms and ivory beads, and over-ran the continent.

If we really want to understand our earliest ancestors, we need to question the model of the mind we use to investigate them, as Wynn and Coolidge have done. What emerges from their book is that modern humans are the biggest obstacle to understanding these people. We are the point of comparison, and because we are so similar to Neanderthals in terms of anatomy, genetics, brain size and, during the Pleistocene epoch, stone technology, the differences become exaggerated. As a result, the Neanderthal 'brand' suffers.

Nevertheless, Wynn and Coolidge show that Neanderthals had a family focus and almost certainly laughed when someone accidentally trod in the fire. They would have recalled that moment among themselves, sharing in the fun through mime and language.

They list nine Neanderthal personality traits. On the negative side, they read the archaeological and fossil evidence as indicating that Neanderthals were xenophobic, resistant to change and dogmatic — direct, but also laconic and unimaginative. The lack of imagination is shown, for instance, in their unchanging tool designs; wariness and xenophobia are indicated by their high mortality rate and interpersonal violence; and their laconic approach is suggested by the fact that they rarely travelled out of their home territory.

On the plus side, the evidence points to Neanderthals as supremely pragmatic, stoic, risk-tolerant when it came to getting food, and both sympathetic and empathetic, caring for disabled individuals in their communities.

Wynn and Coolidge conclude that today, Neanderthals would be commercial fishermen or mechanics, based on their enormous strength and ability to learn the motor procedures needed. Their capacity for empathy might even have made them competent physicians, the authors say, although a lack of mathematical ability means that they

NATURE.COM
For *Nature*'s Web Focus on Neanderthal DNA:
go.nature.com/fr39q6

would never have been able to graduate from medical school. Neanderthals would also make excellent army grunts, with their high levels of pain tolerance, and would be good tacticians in small combat units. They would never rewrite the tactical manual — although tearing it up, however thick, would not be a problem.

Underpinning this appreciation of Neanderthals are two models of how thinking works: expert and embodied cognition. In expert thinking, working memory is not just a short-term store for verbal information. It is important in the planning and execution of complex tasks such as hunting and making tools, as it retains the information necessary to focus the mind and resist interference.

The hand-held technologies of the Neanderthals lead us to embodied cognition. Neanderthals did not think only with their minds but, like us and other primates, through the senses and emotions of the body as well. The tools they used were, Wynn and Coolidge say, "extensions of perception, and hence extensions of mind". Studies of artisans today indicate that a Neanderthal wielding a tool would have learned to respond flexibly through that tool, which effectively became as much a part of their mind as their brain cells.

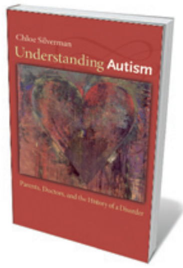
"A Neanderthal's tool effectively became as much a part of their mind as their brain cells."

Embodied cognition is a radical departure in the way the early mind is studied, overturning a long tradition of rational approaches to the mind as a problem-solving machine. In that view, Neanderthals, with their limited technology, did not solve much. However, introducing embodied cognition means that we begin to see many similarities in the emotions they must have felt and the way they dealt with others. The evidence remains the same, but the insights fundamentally change what we believe our distant relatives are capable of.

Read this book for the challenge it poses to the limits of what we can know about our fossil relatives. Delve into its discussion of theory of mind and the ability of humans other than ourselves to think imaginatively about one another's intentions. You will find yourself frequently exclaiming, 'How can they say that?' But I think you will agree that our growing understanding of cognition in deep time could make 'modern human' the real oxymoron. ■

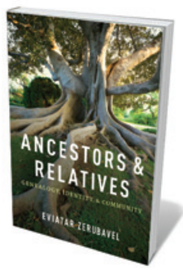
Clive Gamble *is a professor of archaeology at the Centre for the Archaeology of Human Origins, University of Southampton, Southampton S017 1BJ, UK. He is co-editor of* Neanderthals Among Mammoths *(with W. A. Boismier and F. Coward).*
e-mail: clive.gamble@soton.ac.uk

Books in brief

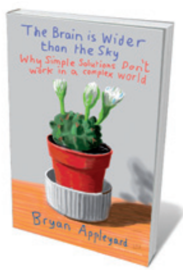


Understanding Autism: Parents, Doctors, and the History of a Disorder
Chloe Silverman PRINCETON UNIVERSITY PRESS *360 pp. $35 (2011)*
Autism remains a contested condition, and given the steep rise in research, diagnosis rates and media coverage, the debate is set to run and run. Science historian Chloe Silverman gives a balanced, sensitive social history of autism that unflinchingly covers many controversial byways. She explores the theory and biomedical advances, and how gene banks, schools and autism organizations have enriched understanding — augmented by parents of children with autism, whose experiences have informed and inspired much research.

Bird on Fire: Lessons from the World's Least Sustainable City
Andrew Ross OXFORD UNIVERSITY PRESS *320 pp. £17.99 (2011)*
These days, Phoenix in Arizona is less eternal firebird than charred turkey. A sprawl of 4.5 million people in an area the size of Taiwan, the metropolis endures a killer combination of scant rainfall, Saharan temperatures and uncontrolled development. Social analyst Andrew Ross interviewed some 200 influential people for his political and environmental analysis, from city planners to eco-activists. Shifting to sustainable 'green democracy', he thinks, will be down to better governance, citing the return of water rights to the Gila River Indian Community, a Native American reservation.



Ancestors & Relatives: Genealogy, Identity, and Community
Eviatar Zerubavel OXFORD UNIVERSITY PRESS *256 pp. £15.99 (2011)*
The issue of relatedness, says sociologist Eviatar Zerubavel, involves more than genealogy. Drawing on genetics, evolutionary biology, anthropology and sociology, Zerubavel looks at kinship and community, the huge role of culture and the "politics of descent" — massaging pedigrees by including distant, impressive ancestors or leaving out recent, mediocre ones. Erudite and amusing, this is also a serious examination of race and ethnicity, asking, for instance, why people with both African and caucasian roots (such as US President Barack Obama) are almost always labelled as black and not biracial.



The Brain is Wider Than the Sky: Why Simple Solutions Don't Work in a Complex World
Bryan Appleyard WEIDENFELD & NICOLSON *289 pp. £20 (2011)*
Is the mind machine-readable? Science writer Bryan Appleyard answers with an emphatic 'no'. Human complexity sits uneasily alongside a life tracked by smart phones and stymied by automated phone operators, he claims. He ponders the often poorly fitting interface between the mind and the new machine age, through a mix of memoirs, history, research and interviews with the likes of Microsoft pioneer Bill Gates. Appleyard concludes that the real issue is not machines, but getting our relationship with them right.

Concrete Planet: The Strange and Fascinating Story of the World's Most Common Man-made Material
Robert Courland PROMETHEUS *416 pp. $26 (2011)*
Forget fossils; the remains of our civilization are more likely to be crushed concrete and oxidized steel, says historian Robert Courland. Concrete may be ubiquitous, but it is a curious substance that repays study. Courland deftly negotiates the chemistry, hydraulics and artistry of concrete in a history that takes in the Neolithic discovery of lime, the 'gold standard' of the Roman Pantheon and the ticking time bomb of today's crumbling concrete infrastructure.

COSMOLOGY

A life in space–time

George Ellis appreciates a Stephen Hawking biography that highlights the epochs of an illustrious career — and the personality behind them.

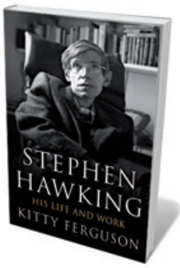
As both a global icon and an innovative theoretical physicist, Stephen Hawking is well served by science writer Kitty Ferguson's fascinating biography. Ferguson explains in accessible terms the major themes that Hawking has explored in his career, and creates a portrait of the private man by drawing on her close personal contact with him and his family.

Hawking — a huge inspiration to disabled people worldwide through his great achievements in the face of motor neurone disease — lives through sheer willpower, yet retains an impish sense of humour and a delight in saying provocative things. For example, when his fame made it difficult for him to be private, he programmed his voice synthesizer to say, "I am often mistaken for Stephen Hawking." He has an incredible determination and adventurousness, epitomized by his taking a zero-*G* flight in 2007 to experience weightlessness.

Born in Oxford in 1942, Hawking found out about his disease when he was just 21, soon after beginning work as a research student at the University of Cambridge, UK. He was able to face this shock through his inner strength, together with strong support from his first wife, Jane Wilde, who married him in the knowledge of his illness — and despite the terrible way he drove while they were out on dates.

Ferguson divides Hawking's career into four epochs. The first, from 1962 to 1973, included his careful technical work on general relativity and cosmology, including his famous critique of the 'action at a distance' alternative theory of gravity by astrophysicists Fred Hoyle and Jayant Narlikar. The highlight was his series of cosmological singularity theorems, developing the work of mathematical physicist Roger Penrose on black-hole singularities. Hawking showed that classical general relativity implies that there was a start to the Universe: a space-time singularity beyond which normal physics would not apply. Hawking also worked on theorems of black-hole geometry and, with his colleagues, established four laws of black-hole thermodynamics.

With this solid work, Hawking built his scientific reputation during this period. But Ferguson's handling of it is thin. She says little about the influential physicists with whom he interacted, including Charles Misner, Robert Geroch, Brandon Carter and John Archibald Wheeler. The book barely mentions Hawking's relation to his research supervisor, Dennis Sciama, who shaped the successful general-relativity research group at Cambridge in the 1960s.



Stephen Hawking: An Unfettered Mind/His Life and Work
KITTY FERGUSON
Palgrave/Bantam: 2011. 320 pp./356 pp. $27/£20

The second epoch, from 1973 to 1979, saw Hawking's adventurous and initially controversial, but later vindicated, work on quantum field theory in a curved space-time. The core is his innovative paper integrating quantum field theory, general relativity and thermodynamics to establish that black holes emit black-body radiation — now known as Hawking radiation. This unexpected result is uniquely his, and is a major achievement that has stood the test of time. He also made important contributions to studying the beginnings of the growth of structure during the inflationary expansion of the early Universe.

I. BERRY/MAGNUM PHOTOS

Stephen Hawking in 1986, a time when his ideas became more speculative, sparking much debate.

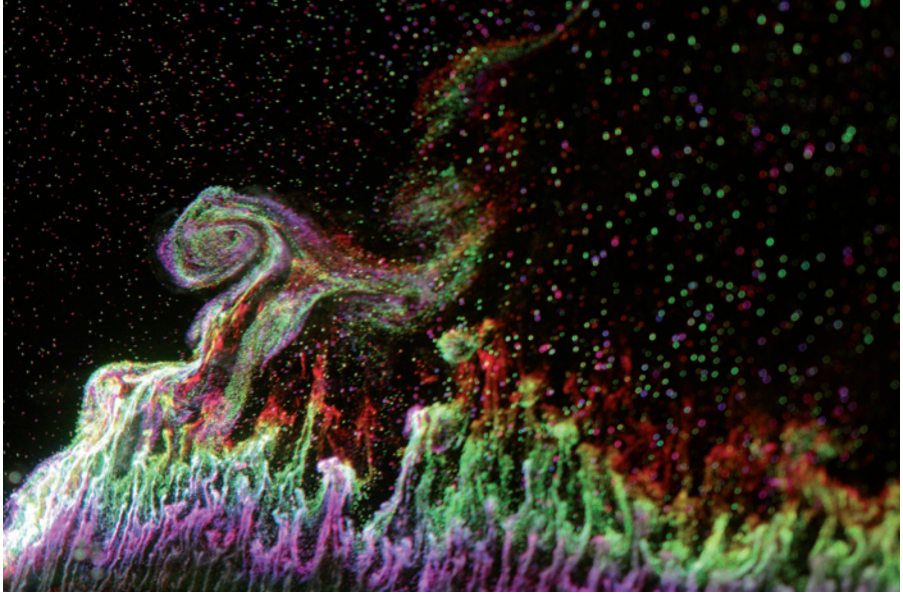
The third era, from 1980 on, was more speculative. Hawking tested big ideas in a creative way, causing much interest and stimulating much activity. But he did not achieve the same level of acceptance for these ideas in the scientific community as he had earlier. They include the 'no-boundary' idea that hypothesizes that the Universe would start without a singularity in a domain where only space existed, as well as his proposals for space-time wormholes. He also proposed that quantum information that has fallen into a black hole is lost at the end of its lifetime. This is a challenging claim for quantum orthodoxy, and has led to much debate.

The fourth period sees Hawking's emergence as a globally admired public figure. He has written an array of popular books, starting in 1988 with the legendary *A Brief History of Time* (Bantam), and has given numerous interviews and talks involving an extraordinary amount of worldwide travel. Hawking's latest book, *The Grand Design* (Bantam, 2010), co-authored with physicist and writer Leonard Mlodinow, considers whether the 'multiverse' concept might explain why the Universe appears the way it does, and why it harbours life (for a review see *Nature* **467,** 657–658; 2010). That solution is still speculative for several reasons, one being that string theory — essential to the version of the concept that he supports — is not experimentally confirmed.

Hawking has recently advocated space travel as a necessity for the survival of the human race, and discussed life elsewhere in the Universe. He suggests that life may be common but that intelligent life is rare. Ferguson goes into Hawking's public pronouncements on issues outside physics — such as his reductionist views on the way the mind works and confusingly varied statements on religion — which can be viewed with caution given that they lie outside his area of expertise.

Ferguson's sympathetic and informed take on an individual who has enriched human knowledge against the odds is an excellent summing-up, as Hawking approaches his 70th birthday, of his unique and creative contribution to both science and humanity. ■

George Ellis *is professor emeritus of applied mathematics at the University of Cape Town, Rondebosch 7701, South Africa.*
e-mail: george.ellis@uct.ac.za



E. DOMNITCH, D. GELFAND

In *Hydrogeny* at the *Surface Tension* exhibition, an electrode in water yields hydrogen, revealed by a laser.

ENVIRONMENT

In at the deep end

A Dublin exhibition inspires a practical approach to water sustainability, finds **Anthony King**.

Surface Tension: The Future of Water
Science Gallery, Trinity College Dublin. Until 20 January 2012.

Seven billion of us rely on the 1% of fresh water that isn't locked away in, say, the ground or the atmosphere. And that 1% is distributed unevenly across the globe. An exhibition at Dublin's Science Gallery seeks to shake those of us in water-rich developed countries out of our complacency about this precious liquid.

Surface Tension brings together scientists, engineers, artists and designers to contemplate water in its forms from oceans to ice, asking two big questions. Can the planet's natural systems sustain our water use? And should water be managed as a commodity or a public good? Nearly 40 exhibits span themes including the scarcity of drinking water; pollution; the 'virtual' water hidden in production processes; the flow of water through cities and the wider water cycle; and future scenarios. The stories are told with urgency and humour, through sculpture and mechanical paraphernalia that tap, cleanse, chart or measure water.

Near the start of the exhibition is an 'intelligent' water meter from South Africa — an unprepossessing grey box about the size of a large cereal packet. Set to allocate each household 25 litres of free water per person a day, the meter makes a simultaneous statement about resource sustainability and social equity. Since the 1990s, municipal authorities have installed thousands of the meters in poor households, says curator Ralph Borland. But South African civil-society groups, such as the Anti-Privatisation Forum, have fought a long-running court case against the devices, which they see as preventing the common ownership of water.

Those 25 litres stand in stark contrast to the 575 litres used daily by the average US citizen. In Ireland and the United Kingdom, that figure is 150 litres — but increases to 3,400 litres when you account for that used in the production of energy, clothes, food and other objects. Food production uses some 70% of the world's fresh water; for example, through irrigation, industrial processing or watering animals. Fittingly, the gallery cafe offers a menu card totting up water footprints. From farm to plate, a steak swallows up 1,550 litres. And one Americano coffee? About 280 litres.

Artist Colin Hart suggests that the British and Irish may one day resort to less-palatable ways of getting a drink. Extrapolating from the use of 'reclaimed' water for drinking in Singapore, Hart presents a tank of canal water — complete with a running shoe and live minnows — which is filtered and offered to visitors to drink.

Avoid the bottled variety, suggests *Bottled Waste*, an exhibit by artist Hal Watts inviting you to turn a pump handle for three hours to fill a one-litre bottle. This effort represents the five megajoules of energy required to make the bottle, purify the water and ship it: more than 1,000 times the energy required to filter and pump a litre of tap water.

Regions with serious water issues are explored, too. Borland says that the state of play "is not so much to do with the physical availability of water, but arguably more to do with politics and power". *Transboundary Waters* is a map of the Middle East showing how the River Jordan, the waters of which are shared between several countries, has become a stage for both conflict and cooperation. For instance, Israelis use 7.5 times more water than Palestinians do, even though the Israeli population is less than twice the size of the Palestinian one. Some 40% of the world's population lives in lake and river basins that are shared by two or more countries (such as the Mekong and the Congo) according to the United Nations — and those regions could be flashpoints in the future.

Sustainability demands that we tease out new relationships with water. Because we use nearly 40% of our daily quota while bathing, showering or brushing our teeth, the exhibit *WaterWise* invites you to step into 2050 and imagine more-sustainable washing routines backed by technology and altered cultural norms. Two scenarios — derived from brainstorming sessions among researchers, product designers and other stakeholders in Ireland's water sector — show waterless washing solutions and dynamic washing, in which washing behaviour changes according to weather fluctuations. The scenarios are playfully depicted in a parody of the emergency instructions for airline passengers.

But what of the oceans, which hold nearly 98% of the planet's water yet host enormous localized soups of pollution? In *The Sea Chair Project*, a hand-powered water pump called a Nurdler sifts beach sand from the southwest of England to extricate nurdles — two-millimetre-diameter plastic pellets that are the raw material for injection moulding and a big problem in the oceans. The project envisages converting fishing vessels into factory ships to recycle this marine debris into chairs.

The varied and sometimes perplexing exhibits don't all paint bleak future waterscapes. One popular installation was a mechanical grid of tubes that gyrates and bobs while suspended on thin cables. Its movements track those of a data buoy lost and adrift somewhere in the Pacific Ocean. Once moored 380 kilometres southwest of Honolulu by the US National Oceanic and Atmospheric Administration, the missing buoy still collects data on wave intensity and frequency, which are scaled down and reproduced in the gallery.

Surface Tension drives home key messages on many aspects of our twenty-first-century relationship with water. In a world in which local depletion and degradation of supplies are acknowledged issues, awareness of local use is simple pragmatism. Equally important is the bigger picture — how research on water management and sustainability can be both liberated and hemmed in by social, political and economic factors. ■

Anthony King *is a writer based in Dublin.*
e-mail: anthonyjking@gmail.com

Correspondence

Fishery threatens protected ocean

The Ross Sea in the Antarctic is the planet's last pristine ocean area, but it could soon become a victim of the race for natural resources at the poles (*Nature* **478,** 174–177; 2011). The region's absolute protection against fishing is being reconsidered by the New Zealand government.

One reason is the demand for a luxury seafood item, the Antarctic toothfish *Dissostichus mawsoni* — a fishery worth NZ$18 million (US$14 million) a year. However, this fish grows slowly and may not spawn every year, so harvesting would be unsustainable.

The designation of the entire Ross Sea as a Marine Protected Area will be debated in November 2012. New Zealand's probable veto was leaked in an official document made public on 11 October (see go.nature.com/ngtelo). The document reveals that the United States, once supportive of Ross Sea protection, is likely to back the New Zealand veto. This has prompted speculation that the move might encourage New Zealand's support for future US ownership claims over Antarctic territories.

A short-sighted refusal by two wealthy nations to protect the Ross Sea's intact marine ecosystem would deprive scientists of invaluable data because its complex structure would be altered forever.

Polar scientists, backed by oceanographer Sylvia Earle, are opposing fishing activities that could remove key species from the ocean's delicately balanced marine food webs. But so far, science-based advocacy for protecting the entire Ross Sea has been glaringly ignored by politicians.

Amélie Lescroël *University of Rennes and the National Museum of Natural History, UMR7204, Rennes, France.*
amelie.lescroel@univ-rennes1.fr
David Grémillet *Centre for Functional and Evolutionary Ecology, CNRS, UMR5175, Montpellier, France; and PFIAO, DST/NRF Centre of Excellence, University of Cape Town, South Africa.*

Women: sexist fiction is alienating

What a surprise to learn that the talent of women for locating objects while shopping comes not from years of experience of domestic chores while our menfolk go hunting for the latest electronics, but from an innate ability to access "womanspace" in parallel universes (E. Rybicki *Nature* **477,** 626; 2011). Perhaps this explains why our gender is so poorly represented in engineering and the physical sciences — we have been operating under an entirely different set of physical principles.

Joking aside, it is hard to laugh off implications that routine domestic duties involve mysterious rites known only to women, and that only men are reliable observers who can make scientific discoveries.

Rybicki's story reflects the pernicious prejudice that biology inherently limits women's success at the highest levels of government, business and science. In our view, it is distasteful to publish fiction that promulgates such sexist notions, even if it was written tongue-in-cheek. We should instead be encouraging the dissolution of the last bastions of 'manspace'.

Ylaine Gerardin, Tami Lieberman *Harvard University, Cambridge, Massachusetts, USA.*
gerardin@fas.harvard.edu

Women: latent bias harms careers

Ed Rybicki's Futures story describes his own helplessness in the face of everyday obstacles (*Nature* **477,** 626; 2011). Although he sees himself as supportive of women scientists, an unintentional, subconscious bias is implied. Such bias can subvert the career path of women — something our community must get to grips with.

The story places women and men in fundamentally different categories: women are well organized and domestically oriented, whereas men are useless in everyday life but come up with theories about the Universe. It is this subconscious categorization that can hurt women as they climb the academic ladder.

Things are better for female scientists now than they were a few decades ago, as overt sexism is slowly dying out. I am hopeful that subconscious bias will follow. Search committees, for instance, could bring these issues out into the open before interviewing candidates for jobs.

Pieter van Dokkum *Yale University, New Haven, Connecticut, USA.*
pieter.vandokkum@yale.edu

Research council will support excellence

The UK Engineering and Physical Sciences Research Council (EPSRC) met last month to discuss calls for further consultation before announcing changes to its funding strategy (*Nature* **477,** 514; 2011). We wish to correct several misunderstandings that exist within the research community.

Research excellence remains pre-eminent and the Council will continue to support applications that are deemed excellent by peer review. We have introduced "national importance" as an additional criterion, but this will not override research excellence.

In deciding which of the areas within our budget we want to grow, maintain or reduce, there are no research areas that we shall completely withdraw from funding. A proposal based on outstanding research will be funded in any area.

Our strategy for supporting training and fellowships will be more targeted, but it will provide flexibility for both individuals and institutions.

We shall continue to work with our advisory teams, stakeholders and research leaders to ensure that we optimize the deployment of our limited resources.

David Delpy, John Armitt *Engineering and Physical Sciences Research Council, Swindon, UK.*
david.delpy@epsrc.ac.uk

Risk assessment for Brazil's GM bean

Your report on the production of genetically modified (GM) beans in Brazil implies that I am an opponent of genetic engineering (*Nature* **478,** 168, 2011). However, you misrepresent my scientific and professional record.

I have never said or written anything against transgenic crops per se. Neither have I claimed to be an opponent of the transgenic technique. However, I have always insisted, as a former member of the Brazilian National Technical Commission on Biosafety (CTNBio) and in my capacity in other professional positions, on critical risk-assessment studies and on research meeting a minimum standard of scientific quality.

This is because proper dossiers from the technology proponents are never presented to the CTNBio or the scientific community. In the case of the transgenic pinto bean from EMBRAPA, the agriculture ministry's research arm, neither of these requirements was met.

Rubens Onofre Nodari *Federal University of Santa Catarina, Florianópolis, Brazil.*
nodari@cca.ufsc.br

CONTRIBUTIONS

Correspondence may be sent to **correspondence@nature.com** after consulting the author guidelines at **go.nature.com/cmchno**.

Herbert Hauptman

(1917–2011)

Mathematician whose theories reveal the shapes of molecules from scattered X-rays.

The intricate molecular structures that regularly grace the covers of scientific journals — including this one — are all monuments to Herbert Hauptman. Fifty years ago, he pioneered mathematical tools for deducing the configurations of molecules as recorded in the patterns of X-rays scattered by crystals. With the architectures of hundreds of thousands of molecules, including numerous drugs — from vitamins to hormones to antibiotics — now established, his efforts have transformed chemistry and biology. In 1985, he shared the Nobel Prize in Chemistry with Jerome Karle. Hauptman died on 23 October, aged 94.

Born in the Bronx in New York City, Hauptman was interested in mathematics from a young age. He earned a bachelor's degree in the subject from the City College of New York in 1937 and a master's degree from Columbia University in 1939. After serving in the US Navy during the Second World War as a weather forecaster in the South Pacific, he moved in 1947 to the Naval Research Laboratory in Washington DC. Here his fruitful collaboration with Karle began. Both were enrolled in the graduate programme at the University of Maryland in College Park, where Hauptman earned his doctorate in 1955.

The combination of Hauptman's mathematical skill and Karle's physical-chemistry background proved powerful in the emerging field of X-ray crystallography. Some researchers — including James Watson, Francis Crick, John Kendrew and Max Perutz — were deducing the shapes of macromolecules, such as DNA and the proteins myoglobin and haemoglobin, through complementary crystallographic techniques. Hauptman, Karle and others sought ways to automatically convert the scattered X-ray patterns from small molecules into structural information.

At the time the only way to do so was to apply the Patterson methods — rules that worked well for some materials, such as organometallic compounds (containing heavy atoms), but were inadequate for others, notably organic molecules. Molecules larger than about 50 atoms remained a challenge, and many drugs, antibiotics and materials of technological interest were beyond reach altogether.

Hauptman and Karle developed mathematical tools — known as direct methods — to convert X-ray crystallography data to molecular forms. In 1953, they introduced two ideas: the most general approach for the solution of the loss of phase information during measurement in crystallography; and the concept of 'structure invariants', which shows the combination of phases that can be estimated from experiments.

G. J. DEL BEL

In the mid-1960s, Hauptman and Karle's mathematical ideas met the applied skills of Michael Woolfson, another father of direct methods. He combined computer algorithms and direct-methods techniques so that many different trial solutions could be explored to find the correct one. Within a few decades, the development of powerful computers and contributions from younger scientists solved the phase problem for molecules of up to 250 atoms. Today, most crystal structures can be computed within minutes.

In 1970, Hauptman moved to the Medical Foundation of Buffalo in New York (renamed in 1994 the Hauptman–Woodward Medical Research Institute), becoming its research director two years later and president in 1988. He turned his mathematical skill to large molecules. Hauptman extended his direct-methods framework to unravel more complex structures by incorporating extra experimental data from complementary crystallographic techniques. Today these methods can be applied to macromolecules with thousands of atoms.

The impact of Hauptman's work goes beyond pinpointing atoms. By establishing bond distances and angles, the chemical activity of a molecule can also be understood — and so, in a biological context, can its function: for example, DNA's double-helical structure is essential to its ability to replicate. This information, now stored for hundreds of thousands of molecules in data banks, is invaluable for chemists, physicists and biologists, whether for designing high-temperature semiconductors, inventing pharmaceuticals or studying biologically active molecules.

Hauptman had a warm personality and was continuously active in teaching direct methods to students of all nationalities in summer schools. He was patient and diligent in his presentation: he spoke quietly, with long pauses so that students could follow his detailed descriptions more easily. To students he was simply 'Herb', always available for a personal tutorial after the official lectures, both before and after having received the Nobel prize.

I first met Hauptman at such a school in 1970, in Parma, Italy. It was there that I decided to dedicate my scientific activity to this field. In the decades since, Hauptman always encouraged my efforts, even recognizing our similar yet independent approaches in his Nobel lecture. We are all in debt to his great generosity. ■

Carmelo Giacovazzo *is professor of crystallography at Bari University, Istituto di Cristallografia, CNR, Via G. Amendola, 122/o, 70126 Bari, Italy.*
e-mail: carmelo.giacovazzo@ic.cnr.it

AGEING

Generations of longevity

The lifespan of some organisms can be extended by mutations that alter how DNA is packaged in their cells. A study reveals that this effect can last for generations, even in descendants that are genetically normal. SEE ARTICLE P.365

SUSAN E. MANGO

For decades, ageing was considered to be the result of progressive damage culminating in catastrophic breakdown. Yet similar animals can have vastly different lifespans — a barn owl can expect to live less than 8 years, a parrot more than 30 years — suggesting that there are mutable genetic pathways that control longevity. By studying small animals with short lifespans, such as the roundworm *Caenorhabditis elegans*, scientists have identified key pathways that can modulate ageing[1]. One of these involves the COMPASS complex (most probably an H3K4me3 regulatory complex in *C. elegans*), an assembly of proteins that organizes DNA by chemically modifying a DNA-packaging protein called histone H3. Mutations in any of the three COMPASS subunits lead to defects in histone regulation and, in *C. elegans*, to a longer life[2].

On page 365 of this issue, Greer *et al.*[3] show that *C. elegans* worms have a memory of COMPASS mutations, even when they themselves carry normal, unmutated COMPASS components: great-grand-offspring and even great-great-grand-offspring live longer if their ancestors lacked COMPASS. In classical genetics, heritable changes in an organism's phenotype (its characteristics) that arise independently of DNA mutations are termed epigenetic inheritance, and have been a source of active research as scientists search for the causative mechanism. An attractive idea is that COMPASS generates an unknown molecular cue that can be inherited over multiple generations to restrict longevity.

Caenorhabditis elegans is not the only organism to show transgenerational effects in health and longevity. Examples can be found in other animals, including humans, and in plants. One of the earliest reported instances stems from studies in Sweden, where historical records revealed that the nutritional and smoking habits of paternal grandparents could influence their descendants' lifespan[4]. In the lab, transient exposure of rats to a high-sugar/low-protein diet leads to glucose intolerance[5], and exposure of rats to noxious chemicals causes reduced fecundity[6]; these effects can last for generations[7]. Although such transgenerational, epigenetic effects have been documented for multiple species, the mechanism remains a mystery.

Greer *et al.*[3] mated *C. elegans* worms that bore mutations in COMPASS subunits with wild-type males (Fig. 1). The resulting progeny, designated F_1, were heterozygous for the mutation (that is, they inherited one mutant copy and one normal copy of the genes that encode COMPASS subunits) and were long-lived like their mutant mothers. Surprisingly, the F_1 animals gave rise to F_2 progeny, F_3 grand-offspring and F_4 great-grand-offspring that were genetically normal but phenotypically long-lived. The authors observed that this effect reverted at the F_5 generation, when the animals' short lifespan was restored. When Greer *et al.* performed similar experiments to examine the influence of other regulators of histones, or of known modulators of longevity such as insulin receptors, they did not observe any transgenerational effects, indicating that the effect was specific to COMPASS.

The authors found that the long life of COMPASS mutants and their descendants required functional germ cells: F_3 animals that lacked germ cells, or that generated unfertilized

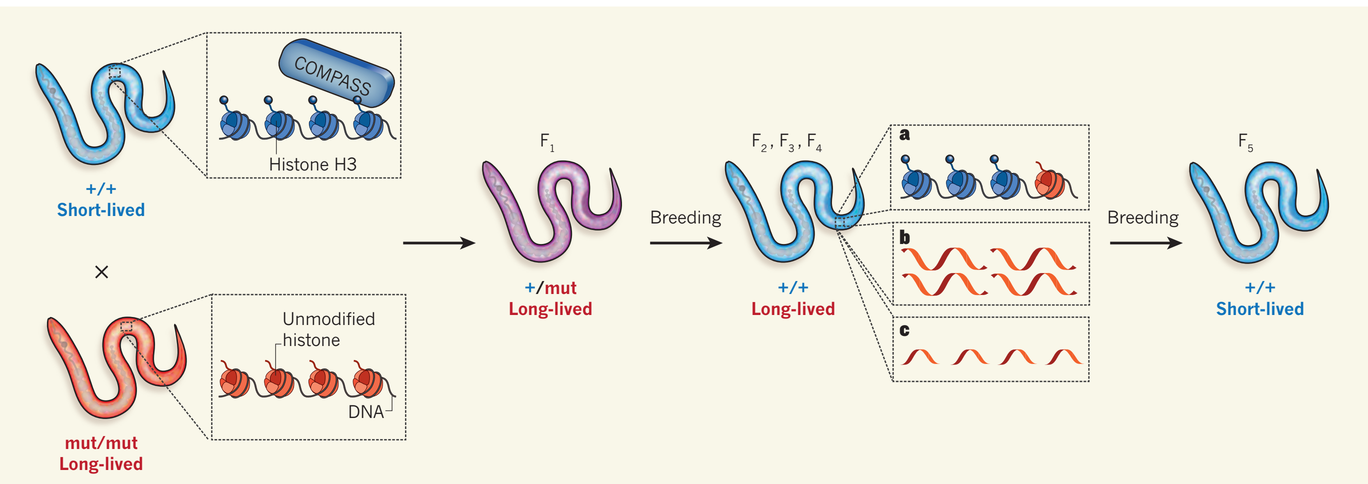


Figure 1 | Transgenerational effects of COMPASS mutations on longevity. The COMPASS protein complex organizes DNA by chemically modifying a DNA-packaging protein called histone H3. Wild-type *Caenorhabditis elegans* worms (blue) are short-lived compared with mutant worms lacking COMPASS (red) . Greer *et al.*[3] crossed wild-type *C. elegans* (which have two copies of the normal COMPASS-encoding gene; +/+) with mutant *C. elegans* (which have two mutant copies of the COMPASS-encoding gene; mut/mut), and obtained long-lived, heterozygous (+/mut) worms (designated as F_1). The next three generations (F_2–F_4) of the worms were genetically normal (+/+), but were still long-lived. The mechanism underpinning the longevity of F_2 to F_4 is unknown, but may reflect: **a**, the presence of unmodified histones at some locations of the genome; **b**, altered expression of protein-encoding messenger RNAs; or **c**, altered expression of small, non-coding RNAs that regulate gene expression. F_5 worms were genetically normal, and not long-lived.

eggs but no sperm, failed to live longer in response to ancestral COMPASS inactivation. The simplest interpretation of this result is that a COMPASS-dependent process occurs in germ cells to control lifespan, and that the flux of germ cells is important for its effects. However, an alternative possibility is that a functional germ line modulates other aspects of worm physiology. For example, it is known that *C. elegans* germ cells undergoing division control fat metabolism in the intestine, and that increased fat metabolism can extend life[8]. A similar non-autonomous effect may account for the germline dependence of COMPASS.

A question central to all transgenerational studies is how transient alterations in the environment or mutation can lead to long-term, multi-generational consequences. One attractive candidate mechanism has been methylation of DNA, a chemical modification that is associated with gene silencing. Most parental DNA methylation is removed in newly fertilized eggs, but some genes may retain methylated DNA over multiple generations, to serve as a transgenerational cue. This effect has been seen in mice, in which a region of DNA that controls coat colour can escape erasure of parental DNA methylation during embryonic development; the remaining degree of DNA methylation produces a range of fur colours[6]. But DNA methylation cannot be the only mechanism for transgenerational signalling, because *C. elegans* lacks DNA methylation altogether. Instead, Greer and colleagues' work[3] implicates histone modifications — specifically, methylation of histone H3 at a particular amino acid (the H3K4me modification).

A crucial issue is whether H3K4me itself is the inherited cue, or whether the modification in parental cells leads to downstream events that generate an epigenetic signal. Previous studies have shown that H3K4me is established in the mother's germ line and inherited by the fertilized egg, where it is retained for at least the first few cell divisions[9]. One possibility, therefore, is that maintenance or interpretation of embryonic histone modifications affects adult worm longevity days later. Greer *et al.*[3] show that their *C. elegans* COMPASS mutants lack H3K4me, and that restoration of H3K4me correlates with a shorter life, suggesting that this histone modification is crucial for controlling longevity. But COMPASS can also function beyond individual genes, to influence the organization of chromosomes — in *C. elegans*, members of COMPASS associate with the dosage-compensation machinery, a complex of proteins that both controls the structure of the X chromosome and attenuates the expression of thousands of X-linked genes[10]. Taken together, the evidence suggests that COMPASS may mediate transgenerational signalling directly, by controlling how genes and chromosomes are organized within cells.

An alternative explanation is that COMPASS and H3K4me activate gene expression, and that the resulting RNA or protein is inherited, rather than the H3K4me mark itself. Greer *et al.* identified genes that were misexpressed — that is, genes that were expressed when they shouldn't have been, and those that weren't expressed when they should have been — in COMPASS mutants, and found that many of the genes coded for proteins associated with longevity, growth or development. In line with the observed transgenerational effects[3], these genes remained misexpressed in the F_4 generation but reverted to normal expression levels in F_5, and the expression of many of the genes was dependent on a functional germ line. An intriguing possibility is that the resulting RNAs are expressed in germline cells, where they could be placed in nascent oocytes (immature egg cells) and passed to the next generation. Although Greer *et al.* focused on messenger RNAs, small, non-coding RNAs (such as siRNAs, miRNAs and piRNAs) are also found in the *C. elegans* germ line and are probably inherited by the embryo. Small RNAs have regulatory roles in silencing gene expression, and could thereby shorten lifespan. Perhaps parental COMPASS is important for producing these non-coding RNAs.

One or a combination of the above explanations may account for the role of COMPASS in lifespan regulation, and it will be fascinating to learn how COMPASS induces transgenerational effects. Future studies will also be able to address whether COMPASS responds to environmental cues such as food, which would link the complex to the types of transgenerational influence that have been described in other animals. Finally, Greer and colleagues' study focused on the role of mothers, but in invertebrates it is clear that both mothers and fathers can signal to their descendants. It will be exciting to learn if COMPASS-dependent cues pass not only through the mother, but also through the father. ■

Susan E. Mango *is in the Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA.*
e-mail: smango@mcb.harvard.edu

1. Kenyon, C. J. *Nature* **464,** 504–512 (2010).
2. Greer, E. L. *et al. Nature* **466,** 383–387 (2010).
3. Greer, E. L. *Nature* **479,** 365–371 (2011).
4. Bygren, L. O., Kaati, G. & Edvinsson, S. *Acta Biotheor.* **49,** 53–59 (2001).
5. Ng, S.-F. *et al. Nature* **467,** 963–966 (2010).
6. Anway, M. D., Cupp, A. S., Uzumcu, M. & Skinner, M. K. *Science* **308,** 1466–1469 (2005).
7. Daxinger, L. & Whitelaw, E. *Genome Res.* **20,** 1623–1628 (2010).
8. Wang, M. C., O'Rourke, E. J. & Ruvkun, G. *Science* **322,** 957–960 (2008).
9. Li, T. & Kelly, W. G. *PLoS Genet.* **7,** e1001349 (2011).
10. Pferdehirt, R. R., Kruesi W. S. & Meyer, B. J. *Genes Dev.* **25,** 499–515 (2011).

QUANTUM PHYSICS

Shaking photons out of the vacuum

The dynamical Casimir effect — the generation of photons out of the quantum vacuum induced by an accelerated body — has been experimentally demonstrated using a superconducting circuit that simulates a moving mirror. **SEE LETTER P.376**

DIEGO A. R. DALVIT

Quantum theory predicts that the vacuum of space is a roiling bath of virtual particles that continuously appear and disappear. These vacuum fluctuations produce measurable phenomena, such as the Casimir effect[1], which arises from the pressure that virtual photons exert on stationary bodies. In 1970, Gerald Moore[2] theorized that bodies in accelerated motion would produce real photons out of quantum vacuum fluctuations — the dynamical Casimir effect. In this issue (page 376), Wilson *et al.*[3] report the first experimental demonstration of the dynamical Casimir effect, using a superconducting circuit that simulates an oscillating mirror.

Accelerated bodies modify quantum vacuum fluctuations, causing emission of photon pairs from the vacuum[4] and dissipation of the bodies' motional energy. The power dissipated in the motion of the body is equal to the total radiated electromagnetic power, as expected according to the law of energy conservation. In its original form, the dynamical Casimir effect was predicted to occur when a single mechanical mirror undergoes accelerated motion in the vacuum. It was then extended to configurations in which the photon production rate is enhanced; for example, in cavities formed by two parallel mirrors, where the position of one of them oscillates with time.

A serious problem for detecting the dynamical Casimir effect induced by moving mechanical systems is that the dissipated energy and the associated radiation are

negligibly small. Among other requirements, motions at nearly the speed of light are necessary. Because of these difficulties, several analogous systems have been proposed for observing the effect, the first being a nonlinear optical medium whose refractive index is rapidly changed with time[5].

In one experiment being pursued[6], the moving mirror is simulated by a semiconducting, layered wall whose conductivity is periodically modulated by an external laser; this set-up closely resembles an actual oscillating mirror. Wilson and colleagues' experiment[3] is based on another proposal[7], and consists of a waveguide terminated at one end by a superconducting quantum interference device (SQUID) — a very sensitive magnetometer. In this approach, a time-dependent magnetic flux threading through the SQUID modifies the electromagnetic field in the waveguide, just as if the SQUID had been replaced by a moving mirror. Because there is no massive body in motion, the effective velocity of the fictitious mirror can be made a substantial fraction of the speed of light.

Even without the dynamical Casimir effect, photons can exist at any finite temperature, and these must be distinguished from motion-induced photons generated from the vacuum. By cooling their apparatus to very low temperatures (less than about 50 millikelvin), Wilson *et al.* prepared their system as close as possible to the vacuum state — the number of thermal photons remaining in such a cold environment is very small. To produce dynamical Casimir photons, the authors 'pumped' the system with a time-varying magnetic flux through the SQUID. They then measured the intensity and frequency of the generated radiation at the open end of the waveguide, as a function of the strength and frequency of the pump field.

Wilson *et al.* detected motion-induced radiation whose broadband microwave energy spectrum was symmetrical at around half the frequency of the oscillating fictitious mirror. The measured spectrum is consistent with that of dynamical Casimir photons, which are generated in pairs whose frequencies add up to the mirror's oscillation frequency. What's more, they found that the measured photon intensity versus pump strength compares reasonably well with theoretical predictions. In addition to observing the creation of real photons, Wilson and colleagues measured photon correlations in the output port of their system. To do this, they split the output photon signal into two separate analysis chains and detected specific correlations. Such correlations are a signature of the quantum nature of the photon-generation process and are another hallmark of the dynamical Casimir effect.

A potential problem with these measurements is that photons might be generated by spurious processes that could mimic the dynamical Casimir effect. Wilson and colleagues considered, and ruled out, a number of such systematic effects. For example, nonlinearities in the electromagnetic properties of the waveguide's substrate and/or in the SQUID electronics could conceivably generate photons in the output port by means of a process known as parametric down-conversion. But the authors emphasize that the pump-power levels used in their experiment are much lower than those needed for such nonlinear processes to occur. Even in the absence of spurious nonlinear mechanisms, motion-induced photons could be seeded, not by quantum vacuum fluctuations, but by uncontrolled noise in the apparatus (for example, thermal noise). However, the authors measured the output photon flux at two temperatures (50 and 250 mK) and were able to verify that the signals are dominated by quantum, and not thermal, fluctuations.

Wilson and colleagues' breakthrough demonstration of the dynamical Casimir effect, together with other ongoing experimental and theoretical efforts, will strongly impact on fundamental physics. They will enable table-top demonstrations of particle creation in an expanding Universe and of black-hole evaporation, among others. ■

Diego A. R. Dalvit *is in the Theoretical Division, MS B213, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.*
e-mail: dalvit@lanl.gov

1. Casimir, H. B. G. *Proc. K. Ned. Akad. Wet.* **51,** 793–795 (1948).
2. Moore, G. T. *J. Math. Phys.* **11,** 2679–2691 (1970).
3. Wilson, C. M. *et al. Nature* **479,** 376–379 (2011).
4. Fulling, S. A. & Davies, P. C. W. *Proc. R. Soc. Lond. A* **348,** 393–414 (1976).
5. Yablonovitch, E. *Phys. Rev. Lett.* **62,** 1742–1745 (1989).
6. Braggio, C. *et al. Europhys. Lett.* **70,** 754–760 (2005).
7. Johansson, J. R., Johansson, G., Wilson, C. M. & Nori, F. *Phys. Rev. Lett.* **103,** 147003 (2009).

GEOPHYSICS

Earth's longest fossil rift-valley system

The origins of the Gamburtsev mountain range, which is hidden beneath Antarctic ice, are a long-standing mystery. Detailed geophysical data from the area form the basis of a comprehensive model that solves the mystery. **SEE LETTER P.388**

JOHN VEEVERS

One-tenth of Earth's crust is masked by the East Antarctic Ice Sheet, and constitutes one of the least understood parts of the planet. Recent work[1,2] reveals that, far from having a flat surface like its former neighbour Australia, this part of the crust is cut into mountain ranges and valleys. The largest of the ranges, the Gamburtsev Subglacial Mountains, is similar in size and shape to the European Alps, but buried beneath kilometres of ice. Its high elevation and jagged topography are intriguing: how could such a topography, which is characteristic of recently formed and uplifted tectonic features, have formed in the interior of an ancient, geologically moribund continent? On page 388 of this issue, Ferraccioli *et al.*[2] present geophysical data that indicate the existence of a 2,500-kilometre-long rift-valley system surrounding an area of anomalously thick crust that is coextensive with the Gamburtsevs, and suggest a model to explain its formation.

The authors' model postulates that the collision of continents about 1 billion years ago produced a high, mountainous topography on thickened crust (see Fig. 4 of the paper[2]). This uplift collapsed under its own weight and was worn away by erosion, whereas the underlying crustal base (known as the root) remained intact. Successive rifting — a process in which a portion of Earth's crust is pulled apart — about 250 million years (Myr) ago, and then again about 100 Myr ago, led to the formation of an extensive rift-valley system with uplifted flanks. The uplifted regions were incised, first by rivers and then (from 34 Myr to 14 Myr ago) by glaciers, to create the steep peaks and valleys of the Gamburtsevs[1]. The youthful topography of the Gamburtsevs was then literally frozen by the growth of the East Antarctic Ice Sheet.

This understanding has come after 60 years of land-based and airborne geophysical surveys. Ice drapes the Gamburtsevs to form the Dome Argus plateau, the highest (4,093 metres) and possibly coldest place in Antarctica. A detailed overland radar survey[1] of the base of the ice at Dome Argus was the first to reveal the underlying alpine topography. An airborne survey of the entire Gamburtsevs, collected as part of Antarctica's Gamburtsev Province Project, now provides details of the mountain range's rugged surface, and constrains speculation about its deep geological structure, as Ferraccioli *et al.*[2] report.

The airborne survey covered a strip

approximately 2,000 km by 1,500 km across East Antarctica, between the South Pole and the terminus of the Lambert glacier. The data reveal that several rifts — the Gamburtsev rifts, the Lambert rift, and some Indian rifts that were continuations of the Lambert rift before continental break-up occurred 130 Myr ago — constituted a rift system that had a similar geometry and length to the East African rift system (Fig. 1). Today, the rift basins in Antarctica host the world's largest subglacial lakes, which resemble the rift lakes of East Africa. The authors' interpretation of the deep geophysical data[2] was guided by comparison with the geophysical signature of the extensive rock exposure in the Prince Charles Mountains, close to the newly identified East Antarctic–Indian rift system.

On the basis of magnetic data acquired from the airborne survey and from satellites, Ferraccioli *et al.*[2] conclude that the surveyed area consists of a mosaic of cratons (rigid blocks) and orogens (regions that have undergone severe structural deformation). The rifts flanking the Gamburtsevs resemble the rifts around the Tanzania craton in the East African rift system. At the Indian end of the East Antarctic–Indian rift system, the rifts are entrenched in orogens that flank the Bastar craton (Fig. 1).

The authors also found that the Gamburtsev province is underlain by crust and lithosphere that are both much thicker than those of the Prince Charles Mountains. Furthermore, the thick root of the Gamburtsev crust has an anomalously high density comparable with that found beneath old orogens that formed when continents collided, and whose dense, lower crustal roots have lost buoyancy on the mantle owing to cooling and metamorphism. Such orogens are exemplified by the approximately 1,800-Myr-old Trans-Hudson collision zone of central North America, and the roughly 300-Myr-old collision zone of Europe and Siberia, in the Urals.

Ferraccioli *et al.*[2] conclude that the crustal root of the Gamburtsev province is 1,800–1,600 Myr old, on the basis of the age of 16 grains of the mineral zircon taken from ice cores in the nearby Vostok province[3]. My group has previously analysed[4,5] about 1,000 zircons found in sands shed from the Gamburtsevs area westward into Dronning Maud Land and northward into the Beaver Lake area during the Permian and Triassic (300–200 Myr ago), and into the Lambert Glacier area in the Cretaceous (about 120 Myr ago), Eocene (about 50 Myr ago) and Quaternary (roughly 2–0 Myr ago). Most of the zircons were 650–500 Myr old, but some were 1,150–850 Myr old. If incorporated into Ferraccioli and colleagues' mosaic model[2], these results[4,5] would suggest that the Gamburtsev province is a craton 1,150–850 Myr old, surrounded by orogens 650–500 Myr old.

Starting from about 300 Myr ago, the Gamburtsev province was regionally elevated ground from which streams drained radially[5]. The Gamburtsev rifts, the Lambert rift and the continuations of the Lambert rift in India initially formed as a result of extension of the crust roughly 250 Myr ago, followed by transtension (sideways pulling apart or wrenching of the crust) about 100 Myr ago[6]. Meanwhile, the lower reaches of the radial drainage in Dronning Maud Land, Beaver Lake and India filled with sediment from the earliest Permian (about 300 Myr ago), as dated by fossils found in the resulting sedimentary rock. The plant fossil *Glossopteris*, for example, is ubiquitous in the Permian sedimentary rocks of Gondwana (the ancient supercontinent that broke up to form modern-day Africa, Antarctica, Australia, India and South America) and is found in pieces of red siltstone in moraine in the area of the southern Lambert glacier[4], which is downflow from the ice summit at Dome Argus (Fig. 1). The siltstone, the only obviously exotic lithology found in the moraines, therefore indicates a Permian age for the oldest sediment in the rifts around the Gamburtsevs.

Ferraccioli *et al.*[2] suggest that the rifting 250 and 100 Myr ago must have enhanced the buoyancy of the crustal root and triggered further uplift of the Gamburtsevs. They propose that the uplift was driven by a combination of the buoyancy of the root and the isostatic response — the readjustment of the level of the crustal root in the underlying mantle — to mechanical and erosional unloading of overlying rock along the rift flank and in the glacial valleys.

Future research should expand the detailed geophysical survey of East Antarctica to cover areas outside the Gamburtsevs, such as the Aurora Subglacial Basin (another region beneath the East Antarctic Ice Sheet), which radar studies have recently found to be characterized by a fjord-like landscape[7]. Surveying the entire bedrock surface and its deep structure will help to pinpoint the oldest ice, which would provide invaluable information about past atmospheres and identify the most suitable bedrock to be sampled by drilling for analyses of its age and composition. In the meantime, Ferraccioli and colleagues' study[2] provides the first comprehensive model to explain how a ragged topography can form in an ice-covered continental interior by modification of the underlying crust and lithosphere. ■

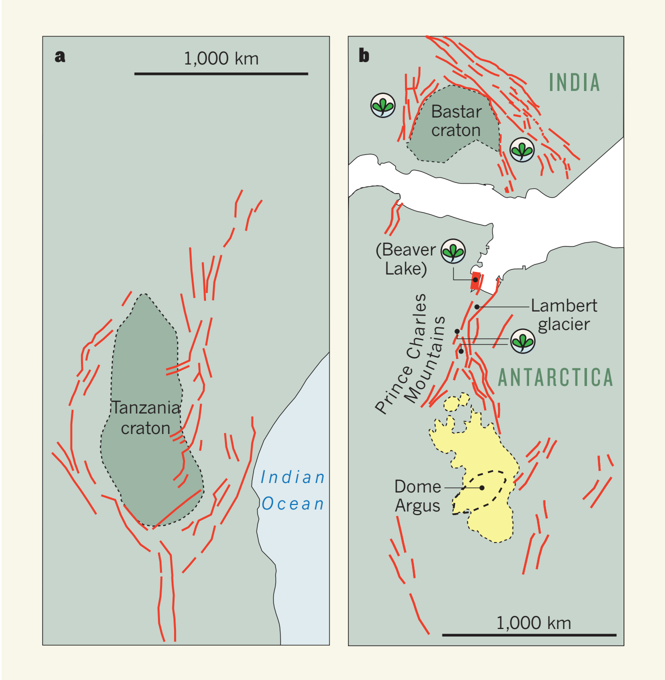


Figure 1 | The largest rift-valley systems. a, The East African rift system[8] consists of a series of rift valleys (red lines) around an ancient, rigid block of rock known as the Tanzania craton. **b**, Ferraccioli *et al.*[2] report that the Gamburtsev Subglacial Mountains (yellow) in Antarctica formed part of a similar rift-valley system 140 million years (Myr) ago, as depicted here, before continental break-up separated India from Antarctica 130 Myr ago[9]. The white region between India and Antarctica is the crustal zone that split when the land masses broke apart. The Indian rifts[10] drape the Bastar Craton. Dome Argus is the summit of the Antarctic Ice Sheet that covers the mountains. Fossils of the plant *Glossopteris* have been found at several locations (leaf symbols), including India[9], Beaver Lake[5] and in moraine on the Lambert glacier[4]. The Indian and Beaver Lake locations correspond to areas that, from 300 Myr ago, received river drainage run-off from the Gamburtsev province; the presence of *Glossopteris* fossils indicates that the corresponding sedimentary rock is about 300–250 Myr old. Moraine on the Lambert glacier was carried by recent ice-drainage run-off from the Gamburtsev province beneath Dome Argus. The Prince Charles Mountains flank the Lambert glacier.

John Veevers *is in the Department of Earth and Planetary Sciences, Macquarie University, Sydney, New South Wales 2109, Australia.*
e-mail: john.veevers@mq.edu.au

1. Bo, S. *et al. Nature* **459,** 690–693 (2009).
2. Ferraccioli, F. *et al. Nature* **479,** 388–392 (2011).
3. Leitchenkov, G. L., Belyatsky, B. V., Rodionov, N. V. & Sergeev, S. A. in *Antarctica: A Keystone in a Changing World — Online Proceedings of the 10th ISAES* (eds Cooper, A. K. *et al.*) paper 014 http://dx.doi.org/10.3133/of2007-1047.srp014 (USGS, 2007).
4. Veevers, J. J., Saeed, A., Pearson, N., Belousova, E. & Kinny, P. D. *Gondwana Res.* **14,** 343–354 (2008).
5. Veevers, J. J., Saeed, A. & O'Brien, P. E.

Sediment. Geol. **211,** 12–32 (2008).
6. Phillips, G. & Läufer, A. L. *Tectonophysics* **471,** 216–224 (2009).
7. Young, D. A. *et al. Nature* **474,** 72–75 (2011).
8. Dawson, J. B. *The Gregory Rift Valley and Neogene — Recent Volcanoes of Northern Tanzania* (Geol. Soc. Lond., 2008).
9. Veevers, J. J. *Gondwana Res.* **16,** 90–108 (2009).
10. Chakraborty, C., Mandal, N. & Ghosh, S. K. *Tectonophysics* **377,** 299–324 (2003).

NEUROSCIENCE

Chemical ecology of pain

The venom of the Texas coral snake causes excruciating pain. The discovery of the venom's pain-inducing component opens up opportunities for studying predator–prey interactions and for pain research. **SEE LETTER P.410**

BALDOMERO M. OLIVERA & RUSSELL W. TEICHERT

Interactions between animals have been investigated primarily through the direct observations of field biologists. But a complementary approach is emerging from an entirely different sector of biology: characterization of molecules that may have evolved in an animal to target another animal, as part of self-defence or predation strategies. On page 410 of this issue, Bohlen *et al.*[1] provide a nice demonstration of this, with their report of the characterization of the pain-inducing component from the venom of the Texas coral snake (*Micrurus tener tener*; Fig. 1). This component, called MitTx, has been honed by natural selection to activate a specific class of signalling macromolecules known as acid-sensing ion channels (ASICs)[2,3]. The characterization of MitTx reveals a previously unknown role in pain signalling for an ASIC subtype known as ASIC1.

Animals are under constant selective pressure to evolve mechanisms that deter potential predators. Venomous snakes are, above all, successful predators, but they can themselves fall victim to predation by other animals. Because they have evolved venom for capturing their own prey, this can also be adapted for defensive purposes.

The compounds evolved by venomous animals for predator deterrence include those that elicit extreme pain. Using an unbiased screen — one that does not rely on mechanistic assumptions — to detect compounds that activate sensory nerve cells, Bohlen *et al.* have identified and characterized just such a pain-inducing compound from the venom of the Texas coral snake. They find that the purified toxin acts through an unexpected molecular mechanism: it elicits pain mainly through ASIC1, which is expressed in sensory neurons. Another ASIC subtype, ASIC3, has been a focus of investigation into pain signalling because its expression is limited mainly to sensory neurons. By contrast, ASIC1 is distributed widely throughout the brain, where it seems to have a variety of functions[4].

In all vertebrates, pH in tissues is rigorously controlled. Significant changes in pH can be caused by acute physiological crises such as massive tissue injury, and may signal the collapse of normal homeostatic mechanisms for controlling pH. All ASICs are activated by low pH — that is, an increase in proton (H^+) concentration — but Bohlen *et al.* find that the ASIC-activating toxin in Texas coral-snake venom is selective for ASIC1 at normal physiological pH. This suggests that ASIC1 channels have a major role in the biological circuitry that triggers an aversive response in the snake's predators.

K. SWITAK/NHPA

Figure 1 | Painful bite. Bohlen *et al.*[1] have identified the pain-inducing component of the Texas coral snake's venom.

Bohlen and colleagues report that two protein components, MitTx-α and MitTx-β, combine to form the functional MitTx complex. MitTx-α is a 'Kunitz' protein, which has a structural scaffold that is found in many inhibitors of protease enzymes (enzymes that degrade cellular proteins). MitTx-β is similar to phospholipase-A_2 enzymes, which are found in a wide variety of snake venoms and have a characteristic structural framework. The phospholipase-A_2-like compound in the Texas coral snake's venom, however, has lost the characteristic activity of these enzymes. The authors find that neither MitTx-α nor MitTx-β alone can activate sensory neurons or ASICs, but that when they are mixed together, the resulting MitTx complex potently and selectively activates ASIC1. The association of a phospholipase-A_2-like protein with a Kunitz-domain protein, and the highly selective activation of ASIC1, are unprecedented.

The authors observe that the MitTx complex is also a weak activator of ASIC2a at physiologically neutral pH. However, it greatly potentiates the activation of ASIC2a by protons as the extracellular pH becomes acidic. This suggests the intriguing possibility that there are undiscovered endogenous ligands that regulate the activity of ASICs. Bohlen *et al.* hypothesize that, if such ligands exist, ASICs may act as 'coincidence detectors' that are activated only in the presence of both extracellular protons and a ligand.

The fact that Texas coral snakes have evolved a potent activator of ASIC channels in their venom that elicits extreme pain raises questions about the interactions of these snakes with predators. Are there structurally and functionally similar molecules in other coral-snake venoms? Is the presence of such a potent pain-inducing molecule coupled to the bright, warning (aposematic) coloration of coral snakes, and do related venomous snakes that don't have aposematic adaptations lack this molecule? And what were the intermediate steps in the evolution of such a highly selective and potent complex of a phospholipase-A_2-like protein with a Kunitz-domain protein? The composition of the complex suggests possible scenarios for how the molecule — and therefore, implicitly, the accompanying biotic interaction — may have evolved.

Characterizing molecules that mediate biotic interactions should thus add another dimension to studies of the ecological relationships between animals that complements field observations. Of course, only field biologists can definitively identify the predators of the Texas coral snake, but knowledge of the molecules used in the snake's defence strategy, combined with field observations, will allow further questions to be answered — for example, does the presence of specific predators trigger an increase in particular toxins in the venom?

Furthermore, the molecules that mediate biotic interactions are often highly selective pharmacological agents. They are therefore useful not only for investigating the ecological relationships between animals, but also as tools for basic research. MitTx is no exception, as it

provides neuroscientists with a new tool for investigating the roles of ASIC1 in pain signalling. This remarkable protein complex will therefore shed light on aspects of both ecology and neuroscience: the defences of coral snakes against predators and the basic mechanisms for sensing pain. ■

Baldomero M. Olivera *and* **Russell W. Teichert** *are in the Department of Biology, University of Utah, Salt Lake City, Utah 84112, USA.*
e-mail: olivera@biology.utah.edu

1. Bohlen, C. J. *et al. Nature* **479,** 410–414 (2011).
2. Waldmann, R., Champigny, G., Bassilana, F., Heurteaux, C. & Lazdunski, M. *Nature* **386,** 173–177 (1997).
3. Waldmann, R. & Lazdunski, M. *Curr. Opin. Neurobiol.* **8,** 418–424 (1998).
4. Wemmie, J. A., Price, M. P. & Welsh, M. J. *Trends Neurosci.* **29,** 578–586 (2006).

QUANTUM INFORMATION

The conundrum of secure positioning

Quantum information has been suggested as a means to prove beyond doubt a person's exact spatial position. But it turns out that all attempts to solve this problem using such an approach are doomed to failure.

GILLES BRASSARD

On 20 July 1969, millions of people held their breath as they watched, live on television, Neil Armstrong set foot on the Moon. Yet Fox Television has reported that a staggering 20% of Americans have had doubts about the Apollo 11 mission. Could it have been a hoax staged by Hollywood studios here on Earth? An article by Buhrman and collaborators[1], published in the *CRYPTO 2011* conference proceedings, comes too late to offer a resolution of this dispute, but studies a related fundamental question: is it possible to devise a scheme by which one person can give definitive proof that he (or one of his agents) is at a specific location?

This 'secure positioning' conundrum has been studied for a number of years. However, Chandran and colleagues proved[2] in 2009 that any possible scheme purporting to solve the problem could be defeated whenever it is based on classical physics. There was great excitement the following year when researchers[3,4] claimed to have obtained perfect solutions to the problem by using the strange and counter-intuitive properties of quantum information. (A US patent had been issued in 2006 for a similar scheme[5], but this did not percolate to the academic world until later.) Unfortunately, these solutions (including the earlier patent) can be foiled by even stranger and more counter-intuitive properties of quantum information[6]. So the race was on within the research community to design a cleverer quantum approach that would be unbreakable. Buhrman and collaborators[1] have just crushed this hope by proving that all attempts to solve this problem are doomed to fall prey to attacks based on quantum teleportation[7], in which the quantum state of a system is instantaneously transferred from one location to another, albeit encrypted until a classical message has been received.

More formally, secure positioning involves a prover and a set of verifiers. The purpose of the verifiers is to ascertain that the prover is at precisely the position he claims to be. We are not concerned about authenticity in the sense that the prover can be replaced by his trusted agent, in which case the agent must be in the purported position. Notwithstanding recent controversy[8], information cannot be transmitted faster than the speed of light, according to Einstein's relativity theory. This makes it easy to ascertain that the prover is no farther than some claimed distance from one fixed verifier, according to the distance-bounding technique[9]: the verifier sends a random signal to the prover, who has to bounce it back immediately. If the signal comes back within one microsecond (10^{-6} s), then the prover (or his agent) cannot be more than 150 metres away because it takes light one microsecond to travel the 300-metre return trip.

It is tempting to infer that the prover can be pinned down to a specific position by asking him to simultaneously demonstrate that he is no more than 150 metres from two verifiers set 300 metres apart, one on either side of the purported position. Unfortunately, this simple-minded approach fails if a cheating prover has two agents, each positioned within 150 metres of one of the verifiers. The purported prover's position can be empty (no one landed on the Moon!), yet both verifiers are simultaneously satisfied, not realizing that they are talking to different agents.

A variety of classical schemes can be designed in attempts to circumvent this predicament, but none will work, according

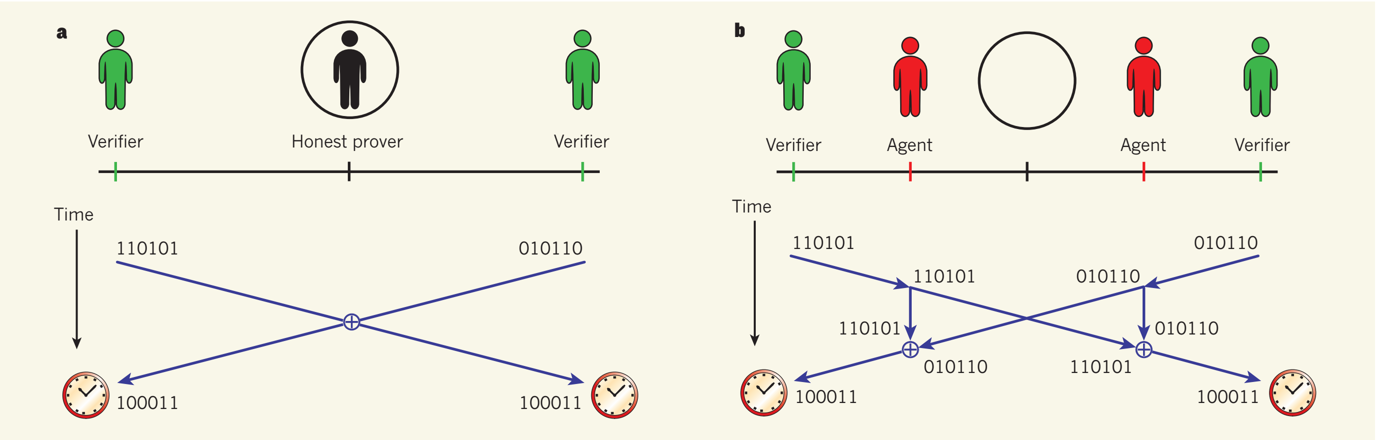


Figure 1 | Cheating classical positioning. Schemes for testing specific spatial positioning involve an honest prover, or a set of dishonest agents, and a set of verifiers. **a**, In this case, each of two verifiers sends a classical string of random bits (zeros and ones) and expects to receive from the prover — promptly after the time it takes light to travel from one verifier to the other through the purported position — something that depends on both their signals, here the result of an operation called bitwise exclusive-or (⊕, where $0 \oplus 0 = 1 \oplus 1 = 0$ and $0 \oplus 1 = 1 \oplus 0 = 1$). **b**, For dishonest agents to produce the same behaviour, even though the purported position is empty, each agent has to keep the input received from his closer verifier and transmit it to the other agent, so that each agent receives both inputs in time to compute and transmit the expected response to his respective verifier. Cheating positioning schemes that are based on sending quantum information is more difficult because of the no-cloning theorem, but Buhrman and colleagues[1] have shown that such schemes can always be defeated by clever use of teleportation-based techniques.

to the impossibility proof of Chandran and collaborators[2]. In general, all such attempts fail because the various agents can copy the information sent to them by some verifiers before forwarding it to other agents (Fig. 1). This is where quantum information comes in. A fundamental property of quantum information is that it cannot be cloned[10,11]. This gives hope that a well-designed quantum scheme could thwart all possible attacks. Indeed, Chandran and colleagues proposed such a scheme[3] in 2010 and 'proved' its correctness. It is informative to explore this doomed attempted solution.

Quantum cryptography[12] is powered by the impossibility of distinguishing between photons whose polarization is horizontal, vertical, at a 45° angle or at a −45° angle. However, it is possible to distinguish perfectly either between horizontal and vertical polarizations, or between 45° angle and −45° angle polarizations, by performing a measurement of either the rectilinear or the diagonal type, respectively. Consider now a prover who claims to be in some specific position and two verifiers situated 150 metres away, one on either side of him. We shall distinguish between the quantum verifier, who sends a polarized photon, and the classical verifier, who sends a bit of information. The verifiers have secretly agreed ahead of time on one of these four polarizations, chosen at random. When positioning has to be demonstrated, the quantum verifier prepares the corresponding photon and sends it to the prover. Simultaneously, the classical verifier sends a message to the prover, telling him which type of measurement to perform on the photon. If properly positioned, the prover can measure the photon's polarization accurately and report it back to both verifiers, who will be satisfied provided they both receive the correct response within one microsecond. To circumvent the risk of a purely lucky correct answer chosen at random by cheating agents, the procedure is repeated several times.

It seems that the correct polarization can be obtained in time only at the purported position because one simultaneously needs the photon and knowledge of the proper measurement type in order to succeed. An agent positioned closer to the quantum verifier would learn the measurement type too late to make the measurement and report on time to the classical verifier. A successful cheating strategy seems to require this agent to both keep the photon and forward it to another agent on the other side of the purported position, which is precisely what the no-cloning theorem forbids. It is remarkable that the two cheating agents can nevertheless fool the verifiers with quantum-teleportation techniques[7] if they share prior entanglement, a fundamental property of quantum mechanics "that enforces its entire departure from classical lines of thought", according to Schrödinger[13]. (See ref. 6 for a detailed description of this attack.)

This teleportation-based approach has been generalized by Buhrman and collaborators[1] to defeat any alleged solution to the secure positioning conundrum, although the cheating agents may need to share a large amount of prior entanglement. More recently, a significantly more efficient general attack has been discovered[14]. On the positive side, Buhrman and colleagues give a provably secure solution under the assumption that the limited technology of would-be cheaters does not allow them to store prior entanglement. Along different lines, a variation on the theme of secure positioning has been proposed, together with a solution and a claimed proof of unconditional security[15].

In my view, the most interesting remaining challenge is to design schemes that are provably secure under the assumption that the cheating agents can share a limited but non-zero amount of prior entanglement. So, did the Americans really land on the Moon in 1969? I would bet that they did, even though it is too late to prove it now. ■

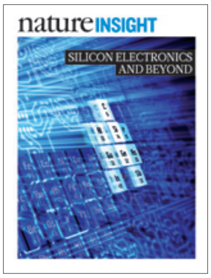
Gilles Brassard *is in the Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, Québec H3C 3J7, Canada.*
e-mail: brassard@iro.umontreal.ca

1. Buhrman, H. *et al.* in *Advances in Cryptology — CRYPTO 2011* (ed. Rogaway, P.) 429–446 (*Lect. Notes Comput. Sci.* Vol. 6841, Springer, 2011).
2. Chandran, N., Goyal, V., Moriarty, R. & Ostrovsky, R. in *Advances in Cryptology — CRYPTO 2009* (ed. Halevi, S.) 391–407 (*Lect. Notes Comput. Sci.* Vol. 5677, Springer, 2009).
3. Chandran, N., Fehr, S., Gelles, R., Goyal, V. & Ostrovsky, R. Preprint at http://arxiv.org/pdf/1005.1750v1 (2010).
4. Malaney, R. A. *Phys. Rev. A* **81,** 042319 (2010).
5. Kent, A. P., Munro, W. J., Spiller, T. P. & Beausoleil, R. G. Tagging systems. US patent 7075438 (2006).
6. Kent, A., Munro, W. J. & Spiller, T. P. *Phys. Rev. A* **84,** 012326 (2011).
7. Bennett, C. H. *et al. Phys. Rev. Lett.* **70,** 1895–1899 (1993).
8. Brumfiel, G. http://dx.doi.org/10.1038/news.2011.554 (2011).
9. Brands, S. & Chaum, D. *Advances in Cryptology — EUROCRYPT '93* (ed. Helleseth, T.) 344–359 (*Lect. Notes Comput. Sci.* Vol. 765, Springer, 1994).
10. Wootters, W. K. & Zurek, W. H. *Nature* **299,** 802–803 (1982).
11. Dieks, D. *Phys. Lett. A* **92,** 271–272 (1982).
12. Bennett, C. H. & Brassard, G. *Proc. Int. Conf. Computers, Systems & Signal Processing*, Bangalore, 175–179 (1984).
13. Schrödinger, E. *Math. Proc. Camb. Phil. Soc.* **31,** 555–563 (1935).
14. Beigi, S. & König, R. Preprint at http://arxiv.org/pdf/1101.1065v1 (2011).
15. Kent, A. *Phys. Rev. A* **84,** 022335 (2011).

nature INSIGHT

SILICON ELECTRONICS AND BEYOND

17 November 2011 / Vol 479 / Issue No 7373



Cover illustration by Nik Spencer (background image: H. Jonsson/ iStockphoto)

Editor, *Nature*
Philip Campbell

Publishing
Nick Campbell

Insights Editor
Karl Ziemelis

Production Editor
Nicola Bailey

Senior Art Editor
Kelly Buckheit Krause

Art Editor
Nik Spencer

Sponsorship
Gerard Preston

Production
Emilia Orviss

Marketing
Elena Woodstock
Hannah Phipps

Editorial Assistant
Hazel Mayhew

Ours is an age of fast-moving information and computer technology. This progress is being driven for a good part by the microelectronics industry, which has been delivering faster and more efficient computers at a remarkably consistent pace. Since the integrated circuit, or silicon chip, was invented in the late 1950s, the number of transistors on a chip has doubled roughly every 18 months — an observation known as Moore's law — so microprocessors can now contain more than two billion transistors.

This achievement is due largely to the design of the classic silicon transistor, which allowed the scaling down of transistors while also improving speed and energy consumption. These triple benefits led to the rise of affordable personal computers in the 1980s and, more recently, to mobile computing technologies such as laptops, smart phones and tablets.

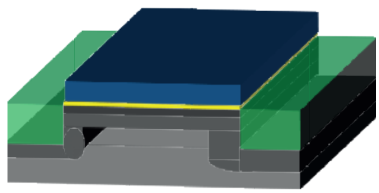
However, transistors cannot scale down indefinitely, and they are now so small that further shrinking would compromise performance. The microelectronics industry is therefore looking beyond the classic silicon transistor to secure the future of a new generation of computers. There is certainly no shortage of new device concepts, but given the existing wide-scale expertise and infrastructure, the best candidates, at least in the short term, are likely to be those that can be integrated within conventional chip technology.

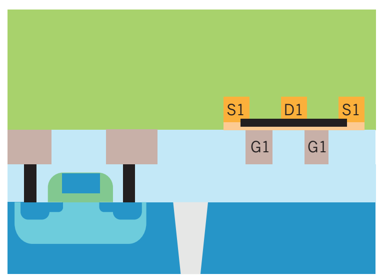
This Insight reviews six promising approaches: some are already at or near the point of manufacturing; others may take another decade (or two). As is imperative for any new technology, the potential environmental impact needs to be assessed, and a Perspective offers some thoughts on such implications.

What is clear is that the silicon transistor is not going to be replaced very soon. However, the integration of new devices with conventional silicon electronics will open up a diverse range of computer applications, from ubiquitous low-power devices to quantum information processors. Silicon-based electronics will continue to drive the information revolution for the foreseeable future.

Liesbeth Venema
Senior Editor

CONTENTS





doi:10.1038/nature10676

Multigate transistors as the future of classical metal–oxide–semiconductor field-effect transistors

Isabelle Ferain[1], Cynthia A. Colinge[1] & Jean-Pierre Colinge[1]

For more than four decades, transistors have been shrinking exponentially in size, and therefore the number of transistors in a single microelectronic chip has been increasing exponentially. Such an increase in packing density was made possible by continually shrinking the metal–oxide–semiconductor field-effect transistor (MOSFET). In the current generation of transistors, the transistor dimensions have shrunk to such an extent that the electrical characteristics of the device can be markedly degraded, making it unlikely that the exponential decrease in transistor size can continue. Recently, however, a new generation of MOSFETs, called multigate transistors, has emerged, and this multigate geometry will allow the continuing enhancement of computer performance into the next decade.

The classic metal–oxide–semiconductor field-effect transistor (MOSFET) is the workhorse of the microelectronics industry. MOSFETs are the building blocks of microprocessors, memory chips and telecommunications microcircuits. A modern microprocessor can contain more than 2 billion MOSFETs, and a 32-gigabyte memory card weighing only 0.5 g contains a staggering 256 billion transistors, which is comparable to the number of stars in the Milky Way. MOSFETs are mainly used as switches in logic microcircuits, although they can fulfil other purposes.

A textbook example of a MOSFET is shown in Fig. 1a. The device consists of two n-type semiconductor regions called the source and the drain, which are separated by a region of p-type semiconductor called the substrate. This description is for an n-channel MOSFET, or NMOS device. A p-type MOSFET, or PMOS device, would have the opposite doping in the source, drain and substrate regions. Typically, the semiconductor is silicon, although other semiconductor materials, with faster charge carriers, are being considered by the microelectronics industry. A thin layer of insulating material such as silicon dioxide covers the region between the source and the drain, and this layer is topped by a metal electrode called the gate. The insulator is referred to as the gate oxide. Under typical bias conditions, the source and the p-type substrate are grounded, and a positive voltage is applied to the drain. Under these conditions, the drain p–n junction is reverse biased, and no current flows between the drain and the substrate. Because the bias across the source p–n junction is zero, there is also no current flowing from the substrate to the source. As a result, there is no current flow between the source and the drain, and the transistor is turned off, playing the part of an open switch. If a large enough positive voltage is applied to the gate, then electrons 'spill out' of the n-type semiconductor source and drain regions, forming an electron-rich layer, called the channel, underneath the gate oxide. The channel forms a continuous electron bridge between the source and the drain, and current can flow between these two electrodes. The transistor is then turned on and behaves as a closed switch. Underneath the electron-rich channel is a region in which holes, which are the charge carriers in p-type semiconductors, have been repulsed and swept away by the positive voltage that has been applied to the gate.

A perfect switch has zero current flow when it is open and zero resistance when it is closed, and it is capable of switching instantly from 'off' to 'on', and vice versa. MOSFETs are, unfortunately, imperfect switches. The off current is not zero; the on current is limited; and switching requires some time. Furthermore, switching does not suddenly occur at a precise gate voltage but takes place gradually over a range of gate voltage values. As transistors are being shrunk in size, their switching behaviour becomes even poorer. One solution is to abandon the planar configuration and to design a gate electrode that is wrapped around several sides of the conducting channel, improving electrostatic control over the channel. Such 'multigate' architectures will allow a further shrinking in size without downgrading transistor performance. In this Review, we discuss why the current generation of MOSFETs has hit a fundamental obstacle for further miniaturization and how the multigate architecture overcomes this obstacle, at least in the short term, and seems set to allow the continuing enhancement of computer performance for another decade.

In pursuit of the perfect switch

Figure 2 illustrates how the drain current that flows through a MOSFET changes as a function of gate voltage for a positive drain voltage of 50 mV. In this example, the on current is 1 mA, and the off current is 50 pA. When the current is plotted on a linear scale, there is no current below a particular gate voltage, called the threshold voltage, which is approximately equal to 0.5 V in the example in Fig. 2. Above the threshold voltage, the current essentially increases linearly with the applied gate bias. When the current is plotted on a logarithmic scale, it is clear that the drain current varies exponentially with the gate voltage below the threshold value and that the off current is not equal to zero. The rate of increase of the current below the threshold voltage is characterized by a parameter called the subthreshold slope (SS), which is defined by the relationship:

$$SS = \frac{dV_G}{d(\log(I_D))}$$

where the logarithm is in base 10, V_G is the gate voltage and I_D is the drain current. The subthreshold slope is expressed in millivolts per decade of current. A typical value for the subthreshold slope of a planar, bulk, single-gate MOSFET is 80 mV decade^{-1}, which means that an 80-mV

[1]Tyndall National Institute, University College Cork, Lee Maltings, Dyke Parade, Cork, Ireland.

increase in the gate voltage brings about a tenfold increase in the drain current (Fig. 2). Thus, to 'switch' the current from its off value (50 pA) to the on state ($I_D = 100\,\mu A$ at the threshold), a swing in gate voltage of

$$80\text{ mV} \times \log\frac{(100\ \mu\text{A})}{50\text{ pA}} = 0.5\text{ V}$$

is required. It can be shown that

$$SS = \frac{k_B T}{q}(\ln(10))\,n$$

where k_B is the Boltzmann constant, T is the temperature in Kelvin, q is the charge of an electron (taking the absolute value because the charge of an electron is negative), ln(10) is the natural logarithm of 10 and n is the body factor. The body factor represents the efficiency, or rather the inefficiency, with which the gate voltage electrostatically controls the channel region. The body factor is proportional to the change in gate voltage with a change in channel potential (Φ_{CH}): that is,

$$n = \frac{dV_G}{d\Phi_{CH}}$$

In the best possible case, if the electrostatic coupling between the gate and the channel region is 100% effective, then

$$n = \frac{dV_G}{d\Phi_{CH}} = 1 \quad \text{and} \quad \frac{k_B T}{q}\ln(10) = 59.6\text{ mV decade}^{-1}$$

at room temperature ($T = 300$ K). In practice, the gate control of the channel region is not perfect because of the electrostatic coupling between the channel and the substrate through the depletion layer. As a result, n typically has a value between 1.2 and 1.5 in bulk MOSFETs, which results in SS values of 70–90 mV decade^{-1}. It is impossible, for thermodynamic reasons, to reduce SS to below 59.6 mV decade^{-1} at room temperature in classical MOSFETs; the best that can be hoped for is to approach this limit as close as possible. The 59.6 mV decade^{-1} barrier can be breached using impact ionization effects[1,2] or quantum-tunnelling effects[3,4] and with special ferroelectric gate materials[5], but none of these techniques has been proven to be reliable or reproducible enough for industrial applications as yet. The lack of scalability of SS is a fundamental limit of MOSFETs and is sometimes referred to as the Boltzmann tyranny[5].

Moore's law

In 1965, Gordon Moore published a classic paper[6] in which he predicted that the density of transistors on a chip would double every 18 months. Even though it is empirical and based on only six years' data (between 1959 when the first silicon integrated circuit was fabricated and 1965), Moore's law has held remarkably well for the past 45 years (Fig. 3).

It is clear that reducing the size of transistors will allow an increase in their density on a chip, which, for a constant chip size, will increase the functionality of the circuit. But there are other incentives for making transistors smaller. Doubling the density of transistors on a chip implies reducing the chip's linear dimensions, such as length and width, by a scaling factor equal to $\sqrt{2}$. This scaling factor is usually represented by κ. In 1974, Dennard and co-workers published a seminal paper in which they demonstrated the benefits of scaling[7]. Based on the assumption of maintaining a constant electric field inside the transistor, Dennard and colleagues demonstrated that scaling the device by a factor κ increases the switching speed by κ, reduces the power dissipation by κ^2 and improves the power-delay product by κ^3. It is worthwhile noting that this scaling law implies a reduction in the supply voltage by κ, as well as a reduction in the threshold voltage by the same factor κ. The latter has not been achieved because of the impossibility of scaling the subthreshold slope to achieve values of less than 59.6 mV decade^{-1}.

Dennard's scaling law was followed by the semiconductor industry until approximately 2005. These years of conventional scaling are now over, and the improvement in performance owing to scaling, at least in terms of microprocessor clock frequency, has reached saturation. This plateau is caused by 'short-channel effects', which arise when the distance separating the source from the drain becomes very small. As can be seen in Fig. 3, the gate length in the current generation of microprocessors is close to 30 nm. In practice, the distance between the source and the drain is approximately 50% shorter than the gate electrode, yielding an effective channel length of only 15 nm. It is predicted that the distance between the source and the drain will be of the order of 5 nm in 2015, which is only 10 times the size of the lattice parameter of a silicon crystal. It is expected that short-channel effects will become even more prominent as devices are scaled down.

Short-channel effects

Short-channel effects result from the sharing of the electrical charges in the channel region between the gate, on the one hand, and the source and the drain, on the other hand. The source and drain junctions create depletion regions that penetrate the channel region from both sides of the gate, thus shortening the effective channel length. These depletion regions carry electric fields that penetrate the channel region to a certain distance and 'steal' some of the control of the channel from the gate. When the drain voltage is increased, this penetration is amplified. As a result, the potential in the channel region and the resultant concentration of electrons are no longer controlled solely by the gate electrode but are also influenced by the distance between the source and the drain and by the voltage applied to the drain. There are two observable effects that result from this loss of charge control by the gate: drain-induced barrier lowering (DIBL), which causes the threshold

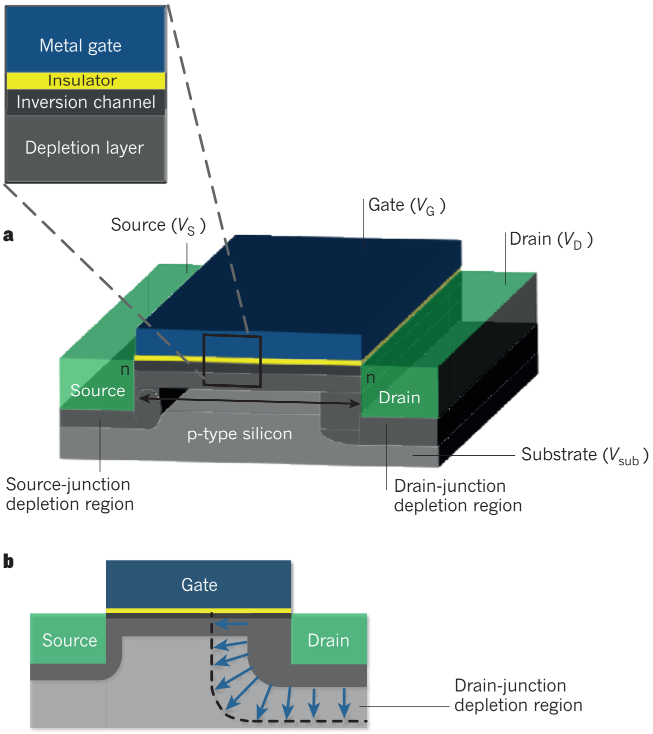


Figure 1 | A schematic view of a classical bulk n-channel MOSFET. a, Two n-type regions called the source and the drain are formed on a p-type substrate. These regions are separated by a distance L, which is called the channel length. A gate stack composed of an insulator and a metal gate electrode is placed above the p-type substrate between the source and the drain. When a positive bias (or positive voltage) is applied to the gate, electrons from the source and the drain are attracted by the gate and form an inversion layer, which is called the channel; the channel connects the source to the drain. Holes in the substrate are repelled by the gate and absorbed by the source and the drain in the vicinity of the junctions, creating hole-starved regions, which are called depletion regions. V_G, gate voltage; V_S, source voltage; V_{sub}, substrate voltage. **b**, The width of the drain-junction depletion region increases (depicted with blue arrows) as the drain voltage (V_D) increases, causing the drain-induced barrier lowering (DIBL) effect.

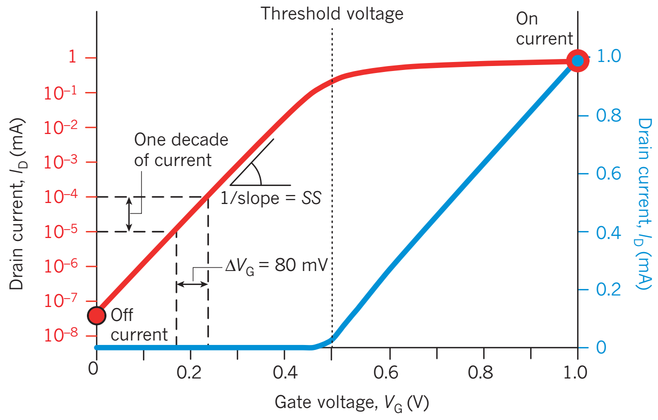


Figure 2 | The drain current as a function of gate voltage in a MOSFET. The two curves show identical data that have been plotted using a linear scale (blue curve; *y* axis, right) and a logarithmic scale (red curve; *y* axis, left). When the gate voltage (V_G) is increased, the number of electrons in the channel increases. This, in turn, increases the current flowing between the source and the drain. The current at the minimum gate voltage (0 V) is the off current, and the current at the maximum gate voltage (1 V in this case) is the on current. Above the threshold voltage (dashed line), the drain current (I_D) increases linearly. Below the threshold voltage, the drain current increases exponentially with the gate voltage. The slope of this exponential increase on a logarithmic scale is called the subthreshold slope (*SS*) and is expressed in millivolts per decade of current. Therefore, if the subthreshold slope is 80 mV decade^{-1}, for example, the gate voltage needs to be increased by 80 mV to increase the subthreshold current by tenfold.

voltage to decrease when the drain voltage increases; and a degradation (that is, an increase) in the subthreshold slope (Fig. 4). The effects are additive and both increase the leakage current of the transistors, constituting a serious impediment to further scaling of MOSFETs. The DIBL effect is graphically illustrated in Fig. 1b: when the drain voltage is increased, the width of the depletion region associated with the drain junction swells and extends laterally underneath the gate.

The loss of switching speed caused by the DIBL effect is given by:

$$\frac{\Delta f}{f} = -\frac{2\mathrm{DIBL}}{V_{DD} - V_{TH}}$$

where f is the maximum operating frequency, V_{DD} is the supply voltage and V_{TH} is the threshold voltage of the transistor. For example, in a circuit operating with a supply voltage of 0.9 V that contains transistors with a threshold voltage of 0.4 V, an increase in DIBL by 50 mV V^{-1} will slow the operating frequency by as much as 20% (ref. 8).

The multigate architecture

In a classical 'bulk' MOSFET, the gate electrode is situated on top of an insulator (an oxide) that covers the channel region of the device between the source and the drain. In such a configuration, the device is planar and essentially two dimensional. Electrostatic control of the channel by the gate is achieved through capacitive coupling between the gate and the channel region, through the gate insulator. The scaling laws require a reduction in the depth of the source and drain regions by the same scaling factor as the gate-length reduction. This reduces short-channel effects by rendering less effective the control of the channel region by the source and the drain. Decreasing the thickness of the gate oxide yields a similar result, by improving the capacitive coupling between the gate and the channel. In addition, replacing silicon dioxide as a gate insulator with other metallic oxides that have a higher dielectric constant can significantly enhance the gate capacitance, which, in turn, yields a higher current. Hafnium oxide and lanthanum lutetium oxide have dielectric constants that are fivefold and eightfold higher than that of silicon dioxide, respectively[9]. Using these materials results in much greater control of the channel by the gate voltage and thus in a reduction in short-channel effects.

The gate's electrostatic control of the channel can also be improved by modifying the shape of the MOSFET. The electrostatics of a long-channel MOSFET are essentially one dimensional. Early textbooks on the physics of semiconductor devices used 'gradual channel approximation', which solves the one-dimensional Poisson equation — the equation that governs the relationship between electric fields and electrical charges — vertically from the gate through the channel and down into the silicon[10]. Short-channel effects, in which electric fields from the source and the drain encroach laterally (horizontally) into the channel region, introduce a second dimension to this problem.

Multigate MOSFETs take advantage of a third dimension to counteract short-channel effects. The term multigate is perhaps not the most appropriate one, as these devices have a single gate electrode. It simply means that the electrode is wrapped around several sides of the channel region. Figure 5 shows several examples of multigate devices, namely fin field-effect transistors (FinFETS), triple-gate (tri-gate) MOSFETS, gate-all-around MOSFETS (in which the gate electrode covers all sides of the channel region) and Π-gate and Ω-gate structures (which are so named because of the shape of their gate electrodes[11,12]). In contrast, the classic MOSFET depicted in Fig. 1 can be called a single-gate transistor.

The multigate transistor portfolio

The first report describing a double-gate MOSFET was published in 1984 (ref. 13). This paper predicted that short-channel characteristics could be improved by using a double-gate architecture instead of the classic, single-gate, approach. The first double-gate MOSFET, the fully depleted lean-channel transistor (DELTA), was fabricated in 1989 and contained a vertically positioned silicon film[14]. Later implementations of vertical-channel, double-gate MOSFETs included FinFETs[15] (Fig. 5a). Tri-gate MOSFETs consist of a thin-film, narrow silicon island with a gate on three sides (Fig. 5b). Implementations of this architecture have included quantum-wire MOSFETs[16] and tri-gate MOSFETs[17]. Improved versions of these feature either a field-induced pseudo-fourth gate, for example in Π-gate MOSFETs[18] (Fig. 5c) and Ω-gate devices[19] (Fig. 5d).

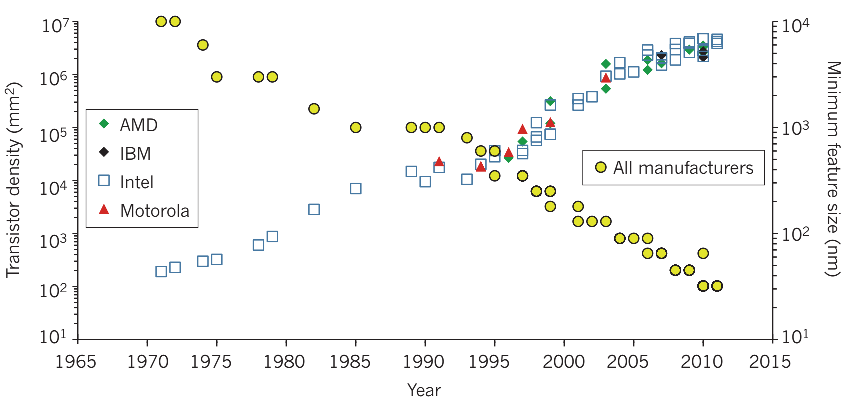


Figure 3 | The evolution of transistor gate length (minimum feature size) and the density of transistors in microprocessors over time. Between 1970 and 2011, the gate length of MOSFETs shrank from 10 μm to 28 nm (yellow circles; *y* axis, right), and the number of transistors per square millimetre increased from 200 to over 1 million (diamonds, triangles and squares show data for the four main microprocessor manufacturers; *y* axis, left). AMD, Advanced Micro Devices; IBM, International Business Machines.

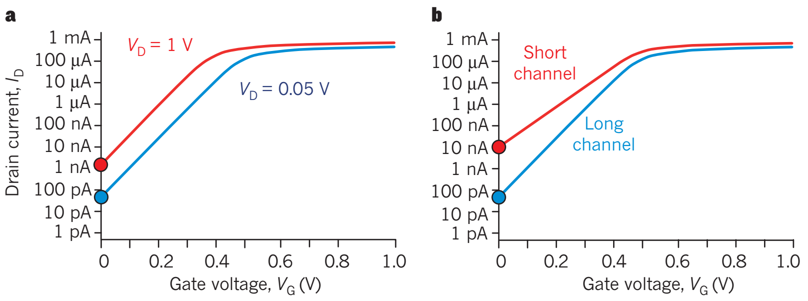


Figure 4 | Illustration of short-channel effects. The blue curves show the relationship between the drain current (I_D) and the gate voltage (V_G) when the drain voltage (V_D) is low (0.05 V), and the red curves show the same relationship when the drain voltage is high (1 V). **a**, The DIBL effect shifts the electrical characteristics of the transistor to the left when the drain voltage is increased. This typically occurs when the device needs to be turned off. **b**, The subthreshold slope increases when the channel length is decreased, which slows down the variation of the current with gate voltage that occurs below the threshold voltage. Both of these effects increase the off currents, which are indicated by the blue and red circles.

The first gate-all-around device was reported in 1990. In this device, the gate electrode is wrapped around all of the sides of the channel region[20] (Fig. 5e). The electrostatic control of the channel by the gate is so efficient in such gate architectures that it is even possible to fabricate a MOSFET device without forming p–n junctions between the source, the channel region and the drain[21]. Such 'junctionless' multigate transistors have the potential to greatly simplify the MOSFET fabrication process at the nanoscale[22]. It is also possible to insert electron-trap layers or nanocrystals in the gate dielectric to create nanowire memory transistors[23,24]. The shortest MOSFET that has been constructed so far has a gate length of 3.8 nm. This MOSFET has a tri-gate structure, a subthreshold slope of 92 mV decade^{-1} and a DIBL of 148 mV V^{-1} (ref. 25). In practice, multigate FETs are usually composed of several parallel nanowire 'fingers' that share a common gate electrode. This configuration allows the current drive of the structure to be increased simply by increasing the number of fingers: for example, in a three-finger Ω-gate MOSFET (Fig. 6).

All of the above devices were made using silicon-on-insulator (SOI) substrates[26,27]. SOI substrates consist of a thin single-crystal silicon layer sitting on top of an insulator, usually silicon dioxide. Multigate FETs can also be made with bulk silicon wafers instead of an SOI substrate. To fabricate such devices, silicon 'fins' are etched on a bulk silicon wafer, and field oxide is deposited to avoid the formation of an inversion layer between the fins. Ion implantation is then used to introduce the desired doping profile to the channel, and the gate stack is deposited. In this architecture, the gate is wrapped around the top and the edges of the silicon fin, creating a tri-gate structure, and the source and the drain are formed by ion implantation (Fig. 5f). Bulk tri-gate devices with a fin width as short as 10 nm have been shown to have good short-channel effect immunity down to the sub-20-nm gate-length regime[28,29]. SOI multigate FETs are easier to fabricate than their bulk counterparts because of the inherent device isolation provided by the buried oxide layer. Bulk substrates, by contrast, are more readily available in large quantities, making large-volume product fabrication much simpler.

Apart from their shape, which takes advantage of three-dimensional space, multigate transistors are similar to conventional devices. Their fabrication process uses the same basic steps as that of any other complementary metal–oxide–semiconductor (CMOS) process, even though some fine-tuning is needed, owing to the different surface topography of the devices. In particular, the 'technology booster' techniques that are used in standard silicon CMOS technology can be applied to multigate FET architectures in a relatively straightforward manner. The most common technology boosters are the use of high-κ gate oxides with metal gates, the use of mechanically strained silicon in the channel (which increases carrier mobility) and the use of silicon epitaxy and metal silicides (which reduces the resistance of the source, the drain and the gate electrodes)[30].

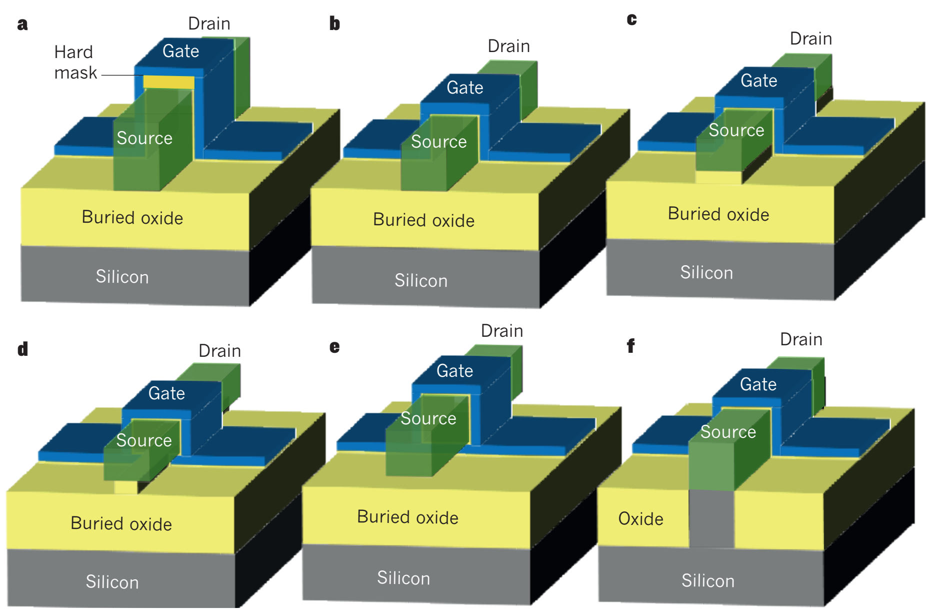


Figure 5 | Types of multigate MOSFET. The different ways in which the gate electrode can be wrapped around the channel region of a transistor are shown. **a**, A silicon-on-insulator (SOI) fin field-effect transistor (FinFET). The 'hard mask' is a thick dielectric that prevents the formation of an inversion channel at the top of the silicon 'fin'. Gate control is exerted on the channel from the lateral sides of the device. **b**, SOI triple-gate (or tri-gate) MOSFET. Gate control is exerted on the channel from three sides of the device (the top, as well as the left and right sides). **c**, SOI Π-gate MOSFET. Gate control is improved over the tri-gate MOSFET shown in **b** because the electric field from the lateral sides of the gate exerts some control on the bottom side of the channel. **d**, SOI Ω-gate MOSFET. Gate control of the bottom of the channel region is better than in the SOI Π-gate MOSFET. The names Π gate and Ω gate reflect the shape of the gates. **e**, SOI gate-all-around MOSFET. Gate control is exerted on the channel from all four sides of the device. **f**, A bulk tri-gate MOSFET. Gate control is exerted on the channel from three sides of the device (the top, the left and the right). In this case, there is no buried oxide underneath the device.

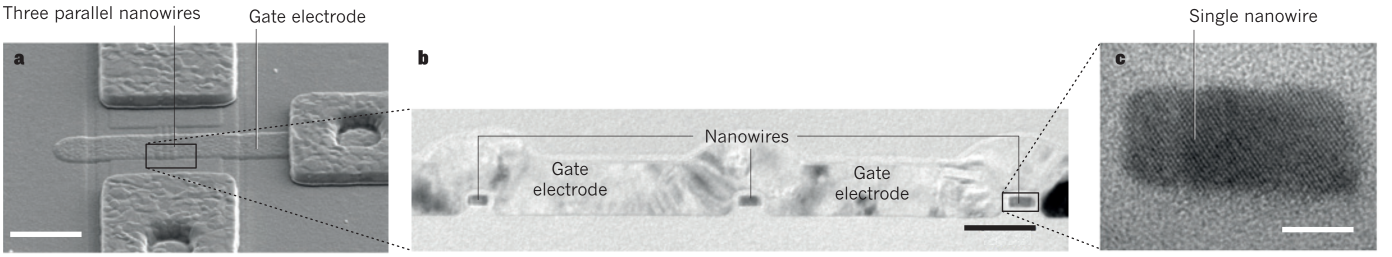


Figure 6 | A multifingered (three-finger) silicon nanowire transistor. a, Scanning electron microscopy image of a device with three parallel nanowires that have a common gate electrode. Scale bar, 5 μm. **b,** Transmission electron microscopy image of the three nanowires and the common polysilicon gate electrode in an Ω-gate electrode configuration. Scale bar, 50 nm. **c,** High-resolution transmission electron microscopy image of a nanowire. Scale bar, 5 nm.

Special devices, such as gated diodes or silicon rectifiers, can be implemented in multigate architectures just as efficiently as they can in planar bulk CMOSs, to protect against electrostatic discharge. Such on-chip devices are connected to the contact pads of integrated circuits and are designed to clamp any static electricity discharge from outside the device that could damage the transistors on the chip[31,32]. The excellent linearity properties and the low value of the body factor of multigate FETs also make them good candidates for mixed-mode and analog circuit applications[33].

Reduction of short-channel effects

Subthreshold swing degradation and DIBL are caused by the encroachment of the electric field line from the source and the drain into the channel region, thereby competing for the available depletion charge and reducing the threshold voltage. The distribution of electrical potential in the channel region of a MOSFET can be derived directly from Maxwell's equation:

$$\nabla \cdot \mathbf{D} = \rho$$

where $\mathbf{D} = \varepsilon\mathbf{E}$ is the electrical displacement field, ε is the permittivity of the material under consideration, $\mathbf{E}$ is the electric field and ρ is the local density of electrical charge.

The three-dimensional Poisson equation shows how the gates compete with the source and the drain for the charge in the channel (Fig. 7):

$$\frac{\mathrm{d}E_x}{\mathrm{d}_x} + \frac{\mathrm{d}E_y}{\mathrm{d}_y} + \frac{\mathrm{d}E_z}{\mathrm{d}_z} = -\frac{\rho}{\varepsilon} = \text{a constant value}$$

For multigate devices, gate control is exerted in the y and z directions and competes with the variation in the electric field in the x direction, which arises from the source and the drain. Because the sum of all of the terms of the Poisson equation is a constant, any increase in the control by the top and bottom gates,

$$\frac{\mathrm{d}E_z}{\mathrm{d}_z}$$

or by the left- and right-hand side gates,

$$\frac{\mathrm{d}E_y}{\mathrm{d}_z}$$

will decrease the penetration of the source and/or drain electric fields into the channel region,

$$\frac{\mathrm{d}E_x}{\mathrm{d}_z}$$

Based on the Poisson equation, it is possible, using a few simplifying assumptions, to calculate a parameter called the natural length, λ, which represents the extension of the electric field lines from the source and the drain into the channel region[11,34]. A device will be free of short-channel effects if the gate is at least sixfold longer than λ. If the transistor has a square cross-section (that is, $t_{\mathrm{Si}} = W_{\mathrm{Si}}$, where t_{Si} is the thickness of the silicon and W_{Si} is the width of the silicon), then the value of the natural length is given by:

$$\lambda_1 = \sqrt{\frac{\varepsilon_{\mathrm{Si}}}{\varepsilon_{\mathrm{ox}}} t_{\mathrm{ox}} t_{\mathrm{Si}}}$$

in a single-gate MOSFET,

$$\lambda_2 = \sqrt{\frac{\varepsilon_{\mathrm{Si}}}{2\varepsilon_{\mathrm{ox}}} t_{\mathrm{ox}} t_{\mathrm{Si}}}$$

in a double-gate MOSFET and

$$\lambda_4 = \sqrt{\frac{\varepsilon_{\mathrm{Si}}}{4\varepsilon_{\mathrm{ox}}} t_{\mathrm{ox}} t_{\mathrm{Si}}}$$

in a gate-all-around (quadruple-gate) MOSFET, where $\varepsilon_{\mathrm{ox}}$ is the electrical permittivity of the gate oxide, $\varepsilon_{\mathrm{Si}}$ is the electrical permittivity of the silicon, t_{ox} is the gate oxide thickness and t_{Si} is the silicon film thickness. These expressions indicate that short-channel effects can be minimized by decreasing the gate oxide thickness, by decreasing the source and drain junction depth (which, in this case, is equal to the silicon film thickness) and by increasing the dielectric constant of the gate oxide material. Increasing this dielectric constant is, from an electrostatic point of view, equivalent to decreasing the 'effective oxide thickness'.

The most interesting information that can be extracted from the calculation of the natural length for the different gate configurations is that

$$\lambda_2 = \sqrt{\frac{1}{2}}\lambda_1 \quad \text{and} \quad \lambda_4 = \sqrt{\frac{1}{4}}\lambda_1$$

It has been shown, using extensive numerical simulations, that the natural length for a tri-gate device is given by

$$\lambda_3 = \sqrt{\frac{1}{3}}\lambda_1 \quad \text{(ref. 35).}$$

The concept of an 'effective gate number', N, can thus be defined, and the generalized relationship for the natural length can be written as follows:

$$\lambda_N = \sqrt{\frac{\varepsilon_{\mathrm{Si}}}{N\varepsilon_{\mathrm{ox}}} t_{\mathrm{ox}} t_{\mathrm{Si}}}$$

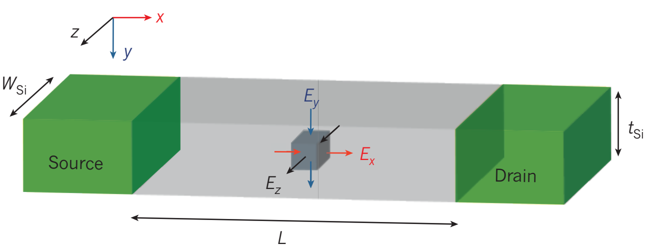


Figure 7 | Competition between the different electric fields for an elemental volume in the channel region. The elemental volume is represented by the small grey cube. The vertical component of the electric field, E_y, arises from the top and bottom gates; the lateral component, E_z, from the side gates; and the longitudinal component, E_x, from the source and drain regions. L, channel length; t_{Si}, thickness of the silicon; W_{Si}, width of the silicon.

which clearly shows the benefits of multigate architectures for reducing short-channel effects. The universality of the effective-gate-number concept is clear from measurements of DIBL for various gate configurations (Fig. 8). The DIBL values for more than 50 numerically simulated devices fall on a single common curve for all gate configurations when the gate length, L, is normalized to the natural length, λ_N. The subthreshold slope shows a similar universal dependence on

$$\frac{L}{\lambda_N} \text{ (ref. 36).}$$

It is worthwhile noting that the natural length of a single-gate MOSFET, and hence short-channel effects such as DIBL, can in theory be decreased by thinning the gate oxide and the silicon film (for SOI devices) or by reducing the junction depth (for bulk devices). In practice, however, these distances are so small that it is not possible to reduce them when the gate length is less than 12–15 nm (ref. 36).

Recent results of *ab initio* device simulations show that gate-all-around transistors with a gate length as short as 3 nm can operate without noticeable short-channel effects[22]. Such devices show how far Moore's law can be stretched.

A proven path to success

Since they were first fabricated, in the early 1990s, multigate transistors have been considered exotic devices that are worthy of academic research but far from industrial mass production. In May 2011, however, semiconductor giant Intel announced its decision to use tri-gate FET devices for its 22-nm technology. This is a clear sign that planar MOSFET scaling is reaching its limits and that short-channel effects can no longer be kept under control using conventional transistor architectures. Even though most of Intel's publications on tri-gate devices deal with SOI devices[37], Intel has chosen to use bulk substrates instead of SOI substrates for its 22-nm tri-gate process. This will reduce costs and minimize the self-heating problems that can arise from thermal insulation of SOI devices from the substrate by the buried oxide. In May this year, the company also demonstrated a computer using a microprocessor named Ivy Bridge, which will be the first high-volume chip to contain tri-gate transistors. The manufacturing of such microprocessors, based on 22-nm three-dimensional transistor technology, is slated for high-volume production readiness by the end of 2011. According to Intel, the tri-gate transistor architecture will make it possible to relentlessly pursue the continuing reduction in chip size that is described by Moore's law and to ensure that the pace of technology advancement that consumers have come to expect will continue for many years.

Other semiconductor giants are also making huge efforts to develop multigate technology for their next-generation products. In December 2010, for example, the Taiwan Semiconductor Manufacturing Company (TSMC) announced a 30-nm FinFET-based process that achieves an on current of 1,400 μA μm^{-1} and an off current of 1.6 nA μm^{-1} at a supply voltage of 1 V (ref. 38).

The compactness of multigate transistors makes it possible to realize extremely compact circuit elements. Static random access memory (SRAM) cells, for instance, are one of the most important building blocks in any logic circuit. Most of the physical space in a modern microprocessor is occupied by memory and not by arithmetic/logic units. It is thus extremely important to shrink memory cells as much as possible to reduce chip area. The excellent performance of multigate SRAM cells was recognized as early as 2004, when high performance, combined with low active and standby power requirements, was demonstrated[39]. The smallest reported SRAM cells were made using multigate transistors, with areas of 0.063 μm^2 and 0.021 μm^2 reported in 2010 and 2011 (refs 40, 41). With these areas, 1.6 billion and 4.8 billion SRAM cells, respectively, fit into a square centimetre. In 2010, TSMC demonstrated SRAM cells that were made using 20-nm FinFETs and operate with a supply voltage as low as 450 mV (ref. 42).

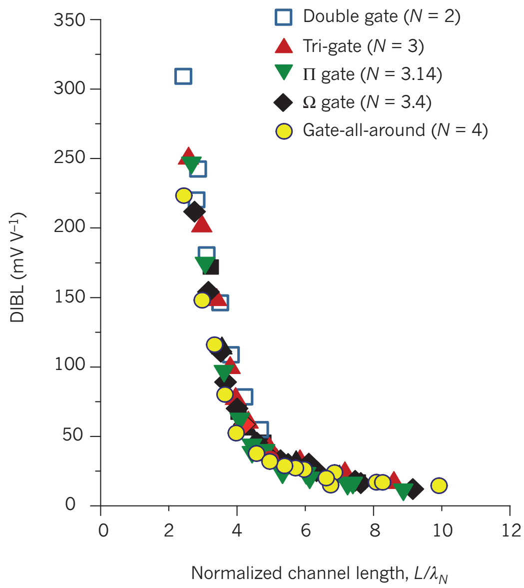


Figure 8 | Variation of the DIBL effect with channel length. The channel length, L, is normalized to the natural length, λ_N. Short-channel effects become noticeable when the channel length is less than six times the natural length. It should be noted that the normalization masks the fact that single-gate devices need to have a channel length of √2, √3 or 2 times larger than that of double-gate, tri-gate or gate-all-around devices, respectively. Data are taken from ref. 35.

Perspectives

This has been a milestone year for the computer chip industry with Intel becoming the first major semiconductor company to affirm its commitment to using multigate transistors in commercial products.

The smallest silicon transistor ever reported is a multigate MOSFET, and *ab initio* simulations show that the use of multigate architectures will allow Moore's law to continue to at least the 3-nm node. This is likely to take the industry at least 20 years to achieve[43].

The multigate architecture lends itself naturally to the fabrication of transistors from a variety of materials, ranging from carbon nanotubes to semiconductor and semimetal nanowires. Combining nanowire transistors with functionalizing radicals and other nanostructures will allow a smooth transition from the realm of microelectronics to the world of nanoelectronics.

Such a transition will, however, pose some design and fabrication control challenges, because quantum-confinement effects begin to appear when the diameter of the nanowire is less than 5 nm. In multigate nanowire transistors, these effects manifest themselves in the form of nonlinearities in the dependence of the drain current on the gate voltage and are highly dependent on the cross-sectional dimensions of the nanowire, requiring control of device fabrication at the atomic level.

Microelectronics have changed our society in ways that no one could have fathomed when Gordon Moore published his visionary paper in 1965. Although it is impossible to predict what kind of electronic gadgets we will be carrying with us in 20 years, we can be sure that the transistors in them will be tiny multigate nanowire devices. ■

1. Armstrong, G. A., Davis, J. R. & Doyle A. Characterization of bipolar snapback and breakdown voltage in thin-film SOI transistors by two-dimensional simulation. *IEEE Trans. Electron Devices* **38,** 328–336 (1991).
2. Moselund, K. E. *et al.* Punch-through impact ionization MOSFET (PIMOS): from device principle to applications. *Solid State Electron.* **52,** 1336–1344 (2008).
3. Zhang, Q., Zhao, W. & Seabaugh, A. Low-subthreshold-swing tunnel transistors. *IEEE Electron Device Lett.* **27,** 297–300 (2006).
4. Afzalian, A., Colinge, J. P. & Flandre, D. Physics of gate modulated resonant tunneling (RT)-FETs: multi-barrier MOSFET for steep slope and high on-current. *Solid State Electron.* **59,** 50–61 (2011).
5. Salahuddin, S. & Datta, S. Use of negative capacitance to provide voltage amplification for low power nanoscale devices. *Nano Lett.* **8,** 405–410 (2008).
6. Moore, G. E. Cramming more components onto integrated circuits. *Electronics* **8,** 114–117 (1965).

7. Dennard, R. H. *et al.* Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE J. Solid-State Circuits* **9,** 256–268 (1974).
8. Skotnicki, T. & Boeuf, T. How can high-mobility channel materials boost or degrade performance in advanced CMOS. *Symp. VLSI Technol.* 153–154 (IEEE, 2010).
9. Engström, O. *et al.* in *Nanoscale CMOS: Innovative Materials, Modeling and Characterization* (ed. Balestra, F.) Ch. 2 (Wiley-ISTE, 2010).
10. Grove, A. S. *Physics and Technology of Semiconductor Devices* Ch. 11 (Wiley, 1967).
11. Colinge, J. P. Multiple-gate SOI MOSFETs. *Solid State Electron.* **48,** 897–905 (2004).
This technical review paper provides a detailed comparison of the efficiency of channel control by the gate with single-gate, double-gate, tri-gate and gate-all-around configurations, and it introduces the concept of natural length and shows its relationship to short-channel effects.
12. Skotnicki, T. *et al.* Innovative materials, devices, and CMOS technologies for low-power mobile multimedia. *IEEE Trans. Electron Devices* **55,** 96–130 (2008).
13. Sekigawa, T. & Hayashi, Y. Calculated threshold-voltage characteristics of an XMOS transistor having an additional bottom gate. *Solid State Electron.* **27,** 827–828 (1984).
14. Hisamoto, D., Kaga, T., Kawamoto, Y. & Takeda, E. A fully depleted lean-channel transistor (DELTA): a novel vertical ultra thin SOI MOSFET. *Tech. Digest IEEE Electron Devices Meet.* 833–836 (IEEE, 1989).
The DELTA transistor was the first multigate transistor, and dynamic random access memory cells based on DELTA devices were reported two years later.
15. Huang, X. *et al.* Sub 50-nm FinFET: PMOS. *Tech. Digest IEEE Electron Devices Meet.* 67–70 (IEEE, 1999).
16. Baie, X., Colinge, J. P., Bayot, V. & Grivei, E. Quantum-wire effects in thin and narrow SOI MOSFETs. *IEEE Int. SOI Conf. Proc.* 66–67 (IEEE, 1995).
17. Doyle, B. S. *et al.* High performance fully-depleted tri-gate CMOS transistors. *IEEE Electron Device Lett.* **24,** 263–265 (2003).
18. Park, J. T., Colinge, J. P. & Diaz, C. H. Pi-gate SOI MOSFET. *IEEE Electron Device Lett.* **22,** 405–406 (2001).
19. Yang, F. L. *et al.* 25 nm CMOS omega FETs. *Tech. Digest IEEE Electron Devices Meet.* 255–258 (IEEE, 2002).
20. Colinge, J. P., Gao, M. H., Romano, A., Maes, H. & Claeys C. Silicon-on-insulator 'gate-all-around device'. *Tech. Digest IEEE Electron Devices Meet.* 595–598 (IEEE, 1990).
21. Colinge, J. P. *et al.* Nanowire transistors without junctions. *Nature Nanotechnol.* **5,** 225–229 (2010).
22. Ansari, L., Feldman, B., Fagas, G., Colinge, J. P. & Greer, J. C. Simulation of junctionless Si nanowire transistors with 3 nm gate length. *Appl. Phys. Lett.* **97,** 062105 (2010).
23. Hofmann, F. *et al.* NVM based on FinFET device structures. *Solid State Electron.* **49,** 1799–1804 (2005).
24. Tang, X. *et al.* Self-aligned SOI nano flash memory device. *Solid State Electron.* **44,** 2259–2264 (2000).
25. Suk, S. D. *et al.* Characteristics of sub 5nm tri-gate nanowire MOSFETs with single and poly Si channels in SOI structure. *Symp. VLSI Technol.* 142–143 (IEEE, 2009).
26. Park, J. T., Colinge, C. A. & Colinge, J. P. Comparison of gate structures for short-channel SOI MOSFETs. *IEEE Int. SOI Conf.* 115–116 (IEEE, 2001).
27. Kuhn, K. J. CMOS transistor scaling past 32nm and implications on variation. *IEEE/SEMI Advanced Semicond. Manuf. Conf.* 241–246 (IEEE, 2010).
28. Okano, K. *et al.* Process integration technology and device characteristics of CMOS FinFET on bulk silicon substrate with sub-10nm fin width and 20nm gate length. *Tech. Digest IEEE Electron Devices Meet.* 725–728 (IEEE, 2005).
29. Cho, H. J. *et al.* Fin width scaling criteria of body-tied FinFET in sub-50 nm regime. *Conf. Digest Device Res. Conf.* 209–210 (IEEE, 2004).
30. Kanemura, T., Izumida, T., Aoki, N. & Kondo, M. Improvement of drive current in bulk-FinFET using full 3D process/device simulations. *Int. Conf. Simulation Semicond. Processes Devices* 131–134 (IEEE, 2006).
31. Cao, S., Chun, J. H., Salman, A. A., Beebe, S. G. & Dutton, R. W. Gate-controlled field-effect diodes and silicon-controlled rectifier for charged-device model ESD protection in advanced SOI technology. *Microelectron. Reliab.* **51,** 756–764 (2011).
32. Thijs, S. *et al.* Advanced ESD power clamp design for SOI FinFET CMOS technology. *Int. Conf. IC Design Technol.* 43–46 (IEEE, 2010).
33. Subramanian, V. *et al.* Planar bulk MOSFETs versus FinFETs: an analog/RF perspective. *IEEE Trans. Electron Devices* **12,** 3071–3079 (2006).
34. Yan, R. H., Ourmazd, A. & Lee, K. F. Scaling the Si MOSFET: from bulk to SOI to bulk. *IEEE Trans. Electron Devices* **39,** 1704–1710 (1992).
35. Lee, C. W. *et al.* Device design guidelines for nano-scale MuGFETs. *Solid State Electron.* **51,** 505–510 (2007).
36. Colinge, J. P. in *FinFETs and Other Multi-Gate Transistors* (ed. Colinge, J. P.) 1–48 (Springer, 2007).
37. Kavalieros, J. *et al.* Tri-gate transistor architecture with high-κ gate dielectrics, metal gates and strain engineering. *Digest Tech. Papers Symp. VLSI Technol.* 50–51 (IEEE, 2006).
38. Yeh, C.-C. *et al.* A low operating power FinFET transistor module featuring scaled gate stack and strain engineering for 32/28nm SoC technology. *IEEE Electron Devices Meet.* 772–775 (IEEE, 2011).
39. Joshi, R. V. *et al.* FinFET SRAM for high-performance low-power applications. *Proc. 34th Eur. Solid-State Device Res. Conf.* 69–72 (IEEE, 2004).
40. Basker, V. *et al.* A 0.063 μm^2 FinFET SRAM cell demonstration with conventional lithography using a novel integration scheme with aggressively scaled fin and gate pitch. *Symp. VLSI Technol.* 19–20 (IEEE, 2010).
41. Guillorn, M. A. *et al.* A 0.021 μm^2 trigate SRAM cell with aggressively scaled gate and contact pitch. *Symp. VLSI Technol.* 64–65 (IEEE, 2011).
42. Wu, C. C. *et al.* High performance 22/20nm FinFET CMOS devices with advanced high-K/metal gate scheme. *IEEE Electron Devices Meet.* 600–603 (IEEE, 2011).
43. ITRS International Technology Working Groups. ITRS 2010 update. *International Road Map for Semiconductors* ⟨http://www.itrs.net/Links/2010ITRS/2010Update/ToPost/2010Tables_ORTC_ITRS.xls⟩ (2010).

Acknowledgements This work was supported by Science Foundation Ireland grants 05/IN/I888, 07/IN.1/I937 and 10/IN.1/I2992, the European project SQWIRE under Grant Agreement No. 257111 and the European Community (EC) Seventh Framework Program through the Network of Excellence Nano-TEC under Contract 257964. We thank N. Petkov and M. Schmidt for the electron microscopy images in Fig. 6.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to J.P.C. (jean-pierre.colinge@tyndall.ie).

doi:10.1038/nature10677

Nanometre-scale electronics with III–V compound semiconductors

Jesús A. del Alamo[1]

For 50 years the exponential rise in the power of electronics has been fuelled by an increase in the density of silicon complementary metal–oxide–semiconductor (CMOS) transistors and improvements to their logic performance. But silicon transistor scaling is now reaching its limits, threatening to end the microelectronics revolution. Attention is turning to a family of materials that is well placed to address this problem: group III–V compound semiconductors. The outstanding electron transport properties of these materials might be central to the development of the first nanometre-scale logic transistors.

The microelectronics revolution might best be characterized by the motto 'smaller is better'. A unique attribute of the silicon metal–oxide–semiconductor field-effect transistor (MOSFET), the workhorse of the industry, is that its logic characteristics improve as its dimensions are reduced[1]. When it comes to logic operations, a transistor behaves as a switch, and its most important qualities, after its footprint, are switching speed and switching energy. Because MOSFETs have decreased in size following a geometrical law, switching speed and transistor density have increased exponentially, while switching energy has decreased in a similar fashion[2,3]. These 'triple dividends' of MOSFET scaling have powered the microelectronics revolution.

Modern logic circuits are based on a pair of transistors with complementary characteristics. They are referred to as n-type and p-type MOSFETs (or simply NMOS and PMOS transistors). Together they are known as complementary metal–oxide–semiconductor (CMOS) transistors and have been the dominant logic family because their simplicity and unique low-power characteristics have allowed the synthesis of very dense circuits.

Recently, MOSFET scaling entered a phase of 'power-constrained scaling' as the power density dissipated by logic chips hit about $100\,\mathrm{W\,cm^{-2}}$ (ref. 4). Power density cannot increase much further without incurring substantial packaging and cooling costs that make these chips impractical for most applications. Continued progress in transistor density will require a reduction in the operating voltage[3–5], but this will compromise switching speed. This problem is partly why the operating voltage for CMOS transistors has bottomed out at around 1 V for some time[3]. Without further reductions, future scaling may not be feasible.

One possible solution is to introduce a new channel material in which charge carriers travel at a much higher velocity than in silicon. This would allow a reduction in voltage without a loss of performance. And this is why attention is turning to III–V compound semiconductors.

The III–V compound semiconductors, such as GaAs, AlAs, InAs, InP and their ternary and quaternary alloys, combine elements in columns III and V of the periodic table. Some III–V compounds have unique optical and electronic properties. Their ability to efficiently emit and detect light means they are often used in lasers, light-emitting diodes and detectors for optical communications, instrumentation and sensing. A few, notably GaAs, InGaAs and InAs, exhibit outstanding electron transport properties. Transistors based on these materials are at the heart of many high-speed and high-frequency electronic systems[6]. In fact, there is a large and mature industry manufacturing III–V integrated circuits in great volumes for applications as diverse as smart phones, cellular base stations, fibre-optic systems, wireless local-area networks, satellite communications, radar, radioastronomy and defence systems. The recent widespread use of handheld devices and their enormous consumption of data has been a boon to the III–V integrated-circuit industry, which is now characterized by highly automated and rigorous large-scale manufacturing, relatively large-area wafers, sophisticated device and circuit design tools, well-established device reliability, and a rich and competitive industrial ecosystem. No other family of materials currently being considered to replace the silicon channel in a MOSFET has such an impressive list of attributes.

Today, III–V CMOS technology is a mainstream part of semiconductor research. Their future role has recently been recognized in the *International Technology Roadmap for Semiconductors*[7].

Here I outline the case for III–V CMOS technology, discuss the most critical problems that remain to be overcome, and summarize recent progress made in the field.

The rationale for using III–V compounds

The case for III–V CMOS technology is often made by drawing attention to the extraordinary electron mobility of certain III–V compounds (Fig. 1). In InGaAs or InAs, the electron mobility is more than 10 times higher than in silicon at a comparable sheet density. The outstanding frequency response of III–V transistors is also frequently invoked. For example, current-gain and power-gain cutoff frequencies of InGaAs-based high-electron-mobility transistors (HEMTs) — a well-established transistor design in its own right — exceed 600 GHz and 1 THz, respectively[8–10]. Impressive as these attributes are, such arguments do not address what really matters for a logic transistor.

A logic transistor operates as a switch that toggles between an 'on' state and an 'off' state. For fast switching, a high on current (I_{ON}) is desired. To limit standby power consumption, the off current (I_{OFF}) must be minimized. It is in terms of I_{ON} and I_{OFF} that the suitability of a transistor for logic should be assessed (for these and other definitions, see Fig. 2 in the Review by Colinge and colleagues[1]).

In an NMOS transistor in saturation, I_{ON} is determined by the product of the sheet electron concentration and the electron injection velocity, v_{inj}, at the 'virtual source'[11], the location on the channel that presents the highest energy barrier in the conduction band. This is the bottleneck to electron flow.

We can learn about the injection velocity of future III–V transistors by

[1]Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

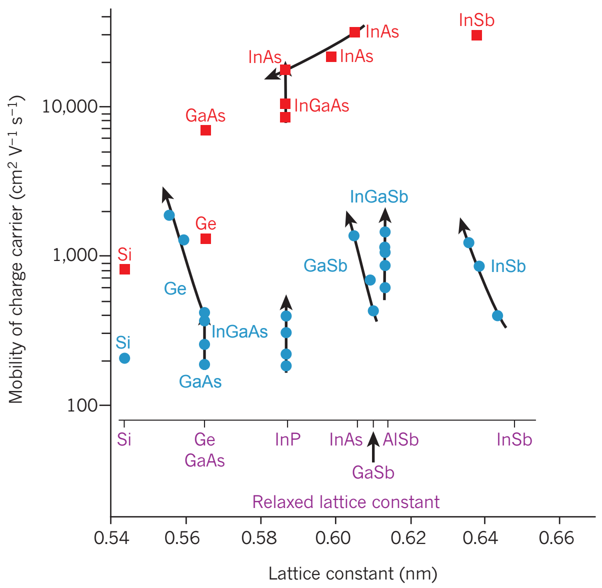


Figure 1 | Electron and hole mobility of group III–V compound semiconductors. The highest room-temperature mobility of electrons (red) and holes (blue) in inversion layers and quantum wells is shown as a function of the actual semiconductor lattice constant (side length of a cubic unit cell of a semiconductor crystal). The mobilities are reported for any sheet carrier concentration. For relaxed layers, under no strain, the lattice constant is its natural one, as shown on the scale. For pseudomorphic layers, which are perfectly strained on a substrate with a different lattice constant, the lattice constant is that of the substrate. As a result, points marked with the same label may appear in different locations in the figure. The impact of biaxial strain is indicated by an arrow representing increasing compressive biaxial strain. There is a wide gap between electron and hole mobilities among III–V compound semiconductors at any lattice constant, and compressive biaxial strain plays a large role in bridging this gap.

examining III–V HEMTs. In this regard, HEMTs provide an excellent model system to study issues of importance in future III–V MOSFETs[12].

Measurements in InGaAs and InAs HEMTs[13] have revealed values for v_{inj} that approach 4×10^7 cm s^{-1} at 0.5 V (Fig. 2). This value of voltage has been selected to compare future technology options because it delivers a sizeable reduction in power dissipation from the present supply of 1 V. In the III–V HEMTs in Fig. 2, v_{inj} is more than twice that of comparable silicon MOSFETs at less than half the voltage[14]. For devices shorter than about 50 nm, the injection velocity becomes independent of gate length. Monte Carlo simulations indicate that electron transport through the channel takes place in a ballistic fashion, that is, with almost no collisions[15]. In this instance, the injection velocity is determined by the band structure of the channel material (Fig. 2), and v_{inj} increases with InAs composition in the channel as a result of a lower electron effective mass[13].

Sheet carrier concentration also affects I_{ON}. Concerns have been expressed about the limitation that a low effective mass imposes on the maximum sheet electron concentration that can be obtained[16]. Recent measurements in InGaAs and InAs HEMTs suggest that the electron effective mass is significantly greater than the bulk value[17]. This is explained by the strong non-parabolicity of the conduction band of these materials, coupled with electron quantization in the thin channel and biaxial compressive stress from lattice mismatch with the InP substrate.

The combination of a high v_{inj} and reasonable channel density of states confers InGaAs and InAs 'quantum-well' FETs with the potential to deliver outstanding I_{ON} at a low supply voltage, V_{DD}, something essential in future CMOS transistors.

But I_{OFF} is just as important as I_{ON}. In quantum-well devices without source and drain junctions, such as HEMTs, I_{OFF} is set by the subthreshold swing, S, which quantifies the sharpness of the drop of the drain current below threshold. In InAs and InGaAs HEMTs, the quantum nature of the channel effectively confines electrons and yields a steep subthreshold behaviour with respect to comparable silicon MOSFETs[18]. The thinner the channel, the closer the subthreshold swing approaches its ideal value of ~60 mV per decade (that is, the current increases by a factor of 10 for every 60-mV increase in gate voltage) at room temperature.

Thinning down the channel is not without drawbacks, however, as scattering tends to increase degrading transport. Measurements of mobility show this. When the thickness of the channel of an InAs HEMT is reduced from 10 nm to 5 nm, the electron mobility degrades from ~13,000 cm^2 V^{-1} s^{-1} to ~10,000 cm^2 V^{-1} s^{-1} (ref. 18). However, the injection velocity of short-gate-length transistors is affected much less[19]. This makes sense because ballistic transport should be expected when transistors are short enough. A thin quantum-well architecture has the potential to scale to very small dimensions.

An important goal of scaling is to maximize I_{ON} while maintaining an acceptable I_{OFF}. When discussing the suitability of different device technologies for logic applications, both values should be considered. A simple way is to refer to the I_{ON} that can be obtained for a set value of I_{OFF} at a certain V_{DD}. This figure can be unambiguously defined in any device with reasonable characteristics even if it does not have the 'correct' threshold voltage, V_T, as is often the case in experimental devices. A standard value for I_{OFF} in high-performance logic devices is 100 nA μm^{-1}. Figure 3 shows the I_{ON} of different devices at this value of I_{OFF} and a V_{DD} of 0.5 V. It includes InAs HEMTs from my own laboratory[20], as well as commercial silicon CMOS transistors scaled to 0.5 V (ref. 12). In addition, projections for future silicon CMOS transistors based on the *International Technology Roadmap for Semiconductors*[7] are also shown. Figure 3 indicates that when appropriately balancing performance and short-channel effects, InAs FETs substantially outperform silicon MOSFETs of similar gate length. The gap is more startling when you consider that the silicon MOSFETs have a source resistance of about 80 Ω μm, compared with 230 Ω μm for the InAs HEMTs. If this shortcoming can be addressed, much better performance can be expected from a future InAs quantum-well FET technology[21].

Underpinning the phenomenal electrical characteristics of III–V FETs is heterostructure growth technology with monoatomic layer precision and an ability to synthesize perfectly specular interfaces. Molecular beam epitaxy (MBE) and, increasingly, metalorganic vapour-phase epitaxy (MOVPE) are at the heart of 'bandgap engineering' in III–V heterostructures. Perhaps the most dramatic testament to these technologies is the electron mobility of 36 million cm^2 V^{-1} s^{-1} obtained at low temperatures in the AlGaAs/GaAs system[22]. This could be the most perfect artificial structure ever made.

The III–V HEMTs have helped make the case for III–V CMOS technology. By themselves, they are not suitable for use in logic because of their high gate leakage current. Nevertheless, HEMTs have provided valuable design features for a future III–V MOSFET, including a junctionless design with a thin, undoped, InAs-rich quantum well that extends under the extrinsic portion of the device, over which is placed raised source and drain regions.

Critical issues

The barriers facing the take-up of a new channel material for CMOS technology are huge. By the time the technology will be ready for deployment, the transistor gate length will need to be shorter than 10 nm. To compound the challenge, a disruptive technology, such as one that incorporates III–V compounds, will need to deliver substantially better performance (at least 30–50% better) than the silicon alternative. It must also promise to deliver more than one future scaled generation. All this must be achieved with cost-effective manufacturing and unprecedented reliability. Before this can happen, several critical problems have to be addressed. These are discussed here.

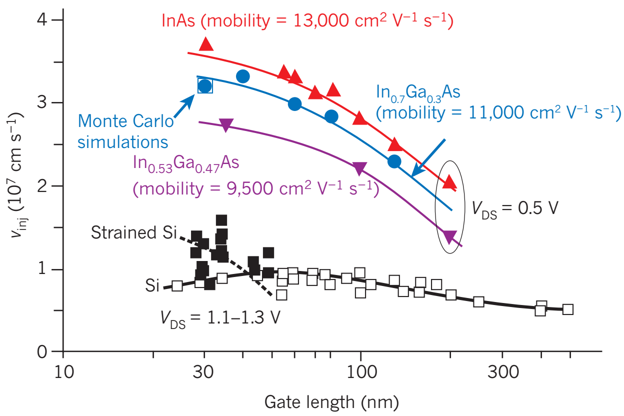


Figure 2 | Electron injection velocity in III–V HEMTs. Electron injection velocity, v_{inj}, is shown for InGaAs and InAs HEMTs with different channel compositions and for silicon MOSFETs as a function of gate length[13,14]. The III–V HEMTs are measured at a drain–source voltage (V_{DS}) of 0.5 V, the silicon MOSFETs at a V_{DS} of 1.1–1.3 V. Despite this discrepancy in voltage, the injection velocity of InGaAs channels is more than twice that of the silicon MOSFETs. The saturation tendency of the injection velocity of InGaAs channels suggests that ballistic (collision-free) transport is occurring; this is confirmed by ballistic Monte Carlo simulations that fall right on the experimental point[15].

The gate stack

At the heart of a MOSFET is the gate stack (Fig. 4). It is composed of a metal gate, a high-permittivity (high-κ) dielectric barrier and the semiconductor channel. It must have a dielectric free of trapped charge and other defects, a smooth interface with few interfacial imperfections, and high stability. One of the miracles of silicon technology is the existence of a native oxide, SiO_2, that meets these requirements. No such native oxides exist for III–V compounds. In fact, exposure of a III–V surface to oxygen results in 'Fermi-level pinning', which is an inability to modulate the electrostatic potential inside the semiconductor[23]. This makes it impossible to use in a MOSFET. In GaAs, the most advanced III–V compound, oxidation creates a rich menu of Ga and As oxides and suboxides, elemental As, As–As dimers and Ga dangling bonds, among other defects[24]. Associated with these is a high density of interface states that prevents the effective modulation of the surface Fermi level[25]. Because of the difficulty of avoiding surface oxidation, early attempts to fabricate GaAs MOSFETs yielded devices with poor performance and low stability[26,27].

In 1995, Ga_2O_3 deposited *in situ* on GaAs was shown to yield an interface that approached the quality of the AlGaAs/GaAs system[28–30]. This led to both n- and p-channel GaAs MOSFETs[31,32] and suggested that dielectric/III–V interfaces with unpinned Fermi levels were indeed possible.

A major step forward was taken in 2003 when a GaAs MOSFET using Al_2O_3 deposited by atomic layer deposition (ALD) was demonstrated[33]. The ALD technique is *ex situ*, robust and highly scalable and is widely used in modern silicon manufacturing, so a high-quality ALD oxide/III–V interface opens the door to manufacturing III–V MOSFETs. This result was unexpected because the starting GaAs surface had been exposed to air. Transmission electron microscopy (TEM) and X-ray photoelectron spectroscopy (XPS) soon showed that during ALD, a kind of 'clean-up effect' takes place in which surface oxides are largely eliminated[34,35]. This happens in the very early stages of ALD[24,36], and subsequent exposure to the ALD oxidant does not regrow the III–V oxides[36]. Using ALD soon led to MOSFET demonstrations on other III–V compounds, such as InGaAs[37], InAs[38] and InP[39].

Progress in the electrical characteristics of III–V MOSFETs accrues from reductions in the interface state density, D_{it}. As their name suggests, interface states are electronic states that arise from disruptions to the ideal bonding structure of a semiconductor at its interface with a dielectric. They affect device operation in several ways. Interface states below the edge of the conduction band increase the subthreshold swing, whereas those inside the conduction band trap electrons. Both effects reduce I_{ON} for a given I_{OFF}. Interface states can also shift the threshold voltage, degrade the channel mobility and be a source of instability.

Insight into the origin of interface states can be gained from density functional theory (DFT) simulations[25,40,41], which allow the construction of detailed bonding models for oxide/III–V interfaces and the computation of the density of states across the band structure. A perfect reconstructed GaAs surface is free of defect states[25]. When the surface is oxidized, interface states appear as a result of As–As bonds, As dangling bonds, Ga vacancies and perhaps Ga–O and As–O bonds[25,40]. Interestingly, GaO-passivated surfaces are seen to be clean of any gap states[42], which is consistent with clear manifestations of unpinned Fermi levels in the Ga_2O_3/GaAs system[43].

According to DFT calculations, several interfacial defects appear when HfO_2 or Al_2O_3, two common ALD high-κ dielectrics, are placed on top of GaAs. Perhaps the most abundant is the As–As dimer, a bonding structure that should not exist in a perfect interface and that is characterized by a near-midgap state[40,41]. In experiments, XPS shows that As–As dimers remain at the interface past the early stages of ALD[24]. A second prominent defect is the Ga dangling bond, which DFT calculations place close to the edge of the conduction band[40,41]. The high density of these defects and their location in the band structure is consistent with experiments and makes the prospect of high-quality GaAs MOSFETs based on ALD oxides problematic[43].

There are several ways to deal with this problem. The first is to engineer the interface through pre-deposition cleaning treatments[44,45], the use of interfacial layers[46,47], the deposition chemistry[48], post-deposition treatments[44,49] or alternative dielectrics[50,51]. The second involves changing the surface crystalline orientation. The most common orientation is (100), but better device results have been reported on the GaAs (111)A surface[52] (this is also the case in InGaAs[53] and InP[54]). This can be explained by DFT as the high-κ/GaAs bonding structure at the interface lacks As–As dimer states[41].

The third approach is to use compounds containing indium. The device characteristics of MOSFETs improve significantly when the InAs mole fraction in the InGaAs channel is increased[55]. InP has also yielded

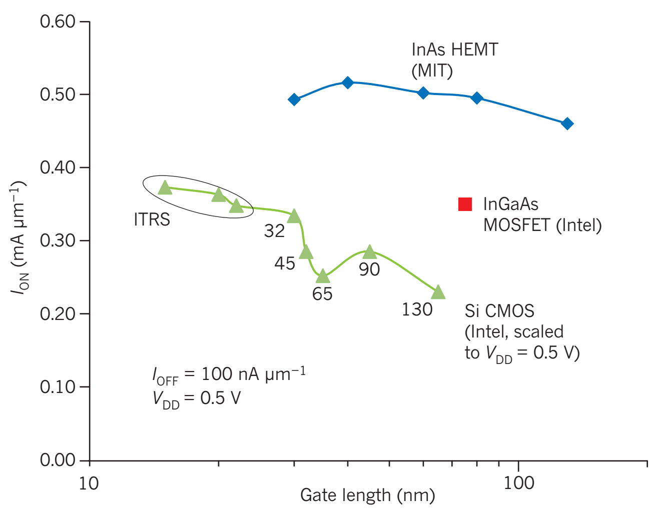


Figure 3 | High 'on' currents in III–V HEMTs. The graph shows how 'on' current, I_{ON}, varies with gate length for InAs HEMTs and silicon MOSFETs at 0.5 V for a fixed 'off' current of 100 nA μm^{-1}. The silicon data correspond to 0.5 V and are obtained from models of published experimental data at higher voltages[12]. The data points labelled ITRS represent projections for future scaling from the *International Technology Roadmap for Semiconductors*[7]. The red square corresponds to an InGaAs MOSFET[59]; the fact that it already exceeds the performance of silicon MOSFETs at 0.5 V is very encouraging. MIT, Massachusetts Institute of Technology. Image modified, with permission from ref. 13.

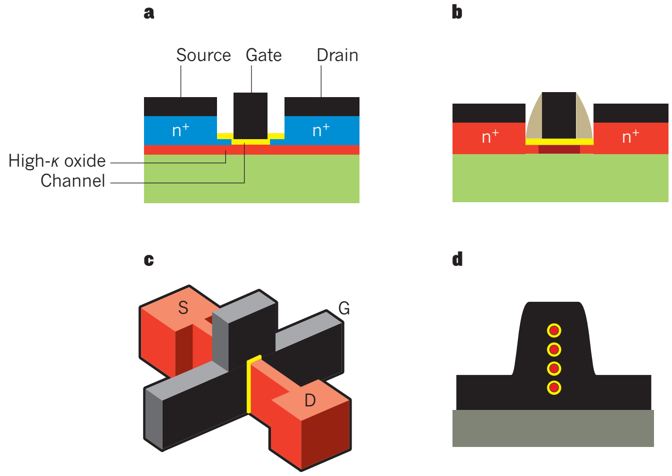


Figure 4 | Possible future MOSFETs using a III–V compound semiconductor channel. **a**, Etched source-and-drain quantum-well MOSFET. **b**, Regrown source-and-drain quantum-well MOSFET. **c**, III–V FinFET, in which the channel charge is electrostatically controlled by a gate that wraps around three sides of a very thin channel. **d**, 'Gate-all-around' nanowire MOSFET, which has an array of very short and thin nanowires with the gate wrapped around them. S, source; G, gate; D, drain.

good results[39]. According to DFT, in these materials the interface states associated with group-V dimers and group-III dangling bonds are predicted to lie well inside the conduction band[41].

A separate consideration in high-κ/III–V MOS structures is the channel mobility. The ideal gate stack from a scaling point of view is a surface-channel device in which the oxide sits directly on top of the channel. The problem is that interface roughness and Coulomb scattering associated with interface states, as well as remote phonon scattering from the high-κ oxide, severely degrade the mobility[56,57]. For InGaAs MOS structures with scaled gate stacks at realistic sheet carrier concentrations, a mobility in excess of 1,000 $cm^2 V^{-1} s^{-1}$ is difficult to obtain[57].

A possible solution to this problem is a buried-channel device. In this approach, a thin wide-bandgap semiconductor is placed between the channel and the oxide. This mitigates the impact of interface roughness and reduces Coulomb scattering from the charged interface and bulk oxide states, as well as remote phonon scattering from oxide phonons. All this yields a higher mobility. However, this approach only goes so far, as the need to manage short-channel effects forces the oxide/semiconductor composite barrier structure to remain very thin. At the moment, this limits electron mobility in InGaAs channels to the 1,000–3,000 $cm^2 V^{-1} s^{-1}$ range[58]. The effect of this level of mobility degradation on the injection velocity in very short devices is not known.

So far, the best III–V MOSFETs have been obtained on InGaAs buried-channel structures equipped with an InP barrier layer using ALD TiSiO as the dielectric[59]. A device with a gate length of 75 nm displays an impressive combination of current drive and subthreshold characteristics, as shown in Fig. 3. This is the only III–V MOSFET I know of with characteristics that entitle it to appear in this figure. Remarkably, its performance exceeds that of state-of-the-art silicon MOSFETs.

Self-aligned transistor design

Using group III–V compounds in CMOS technology only makes sense if they allow further transistor scaling and provide better performance than any of the alternatives. The challenge in making small transistors is twofold. First, it is important to maintain adequate electrostatic integrity. This means that the gate exerts a greater degree of electrostatic control over the electron concentration in the channel than in the drain, which calls for a high geometric aspect ratio for the channel. The second challenge is maintaining low parasitic capacitance and resistance from one part of the structure to another.

Parasitic capacitance is unlikely to be very different in scaled III–V and silicon MOSFETs. Group III–V compounds have slightly higher permittivity than silicon (about 10% higher for GaAs), but this should have only a minor role because parasitic fringe capacitance associated with the gate sidewalls is quickly becoming dominant as devices continue to scale down in size[60].

Parasitic resistance is a significant concern. Future generations of transistors will require a source resistance below 50 Ω μm. State-of-the-art InGaAs HEMTs have a source resistance of about 150–250 Ω μm. The gap is larger than it seems because InGaAs HEMTs feature relatively large contacts (of the order of micrometres). Modelling has shown that when the contact dimensions are appropriately scaled, the contact resistance is more than two orders of magnitude higher than the required value[61]. How is this problem to be solved?

There are several ingredients to the solution. First, the device structure needs to be self-aligned. This means that the contacts are placed without requiring an optical alignment to the gate, as commonly done in HEMTs. Self-alignment allows a gate–contact distance of only a few nanometres. Self-aligned III–V HEMTs have been demonstrated with promising results[61,62].

A very low ohmic contact resistance is also required. Fundamentally, InGaAs should not be at a disadvantage compared with silicon. Doping levels of silicon (the preferred n-type dopant) in InGaAs easily reach the mid-10^{19} cm^{-3} range with an electron mobility that exceeds 1,000 $cm^2 V^{-1} s^{-1}$ at room temperature[63]. This yields a resistivity that beats that of As-doped silicon with doping levels in the mid-10^{20} cm^{-3} range with one-tenth of the mobility. Refractory ohmic contacts to n^+-InGaAs with contact resistivities of around 1–2 Ω $μm^2$ have been demonstrated using different metals[63]. These values are comparable to the best contacts to n^+-Si and are in the range required for fully scaled devices.

The biggest concern about scaled III–V transistors is attaining adequate electrostatic integrity. This refers to the tight control of the channel charge by the gate, a key requirement for sharp subthreshold swing. In a planar quantum-well design, this demands a very thin channel and an extremely thin gate barrier.

Two possible planar quantum-well designs are shown in Fig. 4. Figure 4a shows a device architecture with source and drain regions grown with the original heterostructure and recessed to accommodate a self-aligned gate. In this design, the quantum well extends underneath the source and drain, and high-mobility transport is preserved in the extrinsic device. A second advantage is that the dielectric/III–V interface is formed relatively late in the process, providing substantial process flexibility. The InGaAs MOSFET illustrated in Fig. 3 has a structure like this[59].

A second possible device design is shown in Fig. 4b. In this architecture, the gate stack is formed early in the process. Using the gate as a mask, the channel is etched away from the extrinsic portion of the heterostructure, and then the source and drain regions are grown epitaxially in a self-aligned way. A potential advantage of this approach is the ability to introduce uniaxial strain in the channel. Prototype devices have been fabricated exhibiting promising electrical characteristics[64,65].

Should the planar design fail to meet the requirements, alternative device structures exist. Recently, Intel announced the use of a trigate FET[66] for the 22-nm CMOS generation. The trigate, also known as a FinFET, is in essence a MOSFET in which the channel charge is electrostatically controlled by a gate that wraps around three sides of a very thin channel. This approach yields improved electrostatic control and scalability. Similar devices based on III–V compounds (Fig. 4c) have already been demonstrated with improved short-channel effects over planar designs[67,68]. An important concern is dry-etching damage, which is difficult to anneal in compound semiconductors[69].

Higher electrostatic integrity and scaling potential are expected from nanowire FETs (Fig. 4d). These consist of an array of very short and thin nanowires with the gate wrapped around them. Silicon nanowire

FETs have been studied for some time[70,71] and constitute an alternative CMOS technology in their own right. For III–V compounds, horizontal and vertical nanowire FETs with impressive characteristics have been demonstrated in the InAs system[72,73].

The p-type MOSFET

Both NMOS and PMOS transistors with reasonably matched performance are required for CMOS logic circuits. The PMOS transistors are based on holes and tend to be inferior to NMOS transistors because of their generally lower mobility. Circuit designers have learned to work with a silicon PMOS transistor that has about one-third of the current density of the NMOS transistor. A future III–V CMOS technology should strive for a performance gap that is no worse than this.

Although hole mobility offers limited guidance in the selection of a suitable p-type channel material, the large imbalance between electron and hole mobilities in III–V compounds is a serious problem, as Fig. 1 shows. It depicts the highest electron and hole mobilities at room temperature that have been reported in inversion layers or quantum wells in various semiconductors as a function of the heterostructure lattice constant (the actual lattice constant, as opposed to the relaxed lattice constant). These mobilities typically come from buried-channel heterostructures (such as HEMTs) at relatively low carrier concentrations.

In Fig. 1, not one semiconductor features both an electron mobility above 5,000 $cm^2 V^{-1} s^{-1}$ and a hole mobility above 1,500 $cm^2 V^{-1} s^{-1}$. In most cases the hole mobility is significantly lower than this. There is not even a pair of materials with a similar lattice constant that exhibit mobilities in those ranges. Perhaps the closest is GaAs for the NMOS transistor and Ge for the PMOS transistor, but I have already discussed the difficulty in obtaining high-quality high-κ/GaAs interfaces by ALD.

It seems inevitable then to conclude that a future III–V CMOS technology will feature NMOS and PMOS transistors made of different materials with different lattice constants. This has important implications for their co-integration on a common substrate, as discussed below.

The hole mobility can be increased by introducing compressive biaxial strain. This can be accomplished through pseudomorphic growth on a material with a smaller lattice constant. Figure 1 illustrates this through the arrows, which indicate increasing compressive biaxial strain. The gains can be substantial. A hole mobility approaching 2,000 $cm^2 V^{-1} s^{-1}$ has been reported in compressively strained Ge[74], with around 1,500 $cm^2 V^{-1} s^{-1}$ in InGaSb[75], more than 1,300 $cm^2 V^{-1} s^{-1}$ in GaSb[76], and over 1,200 $cm^2 V^{-1} s^{-1}$ in InSb[77]. These are now the most promising materials for PMOS transistors.

There are other ways of improving hole mobility. In Si and Ge, uniaxial compressive strain also increases the hole mobility[78]. Through uniaxial strain, silicon PMOS transistors have recently narrowed the performance gap with n-type MOSFETs[60]. Uniaxial strain also represents a promising approach for III–V compounds[79–81]. An interesting recent finding is that the superposition of uniaxial strain and biaxial strain yields nonlinear mobility gains[80,82].

There have already been demonstrations of PMOS transistors in GaAs[32], InGaAs[83], GaSb[84] and InGaSb[84]. Most of the issues discussed above for NMOS transistors also apply to PMOS transistors, but their development lags behind. At the moment, it does not seem that any III–V PMOS transistor will have a performance advantage over a Ge device, a technology that is much more mature[85]. For this reason, at the present time, the leading PMOS contender for a III–V CMOS technology is based on Ge.

Co-integration of NMOS and PMOS transistors on silicon

Perhaps progress is most needed in the side-by-side integration of III–V NMOS and PMOS transistors on a silicon substrate. Economics dictates the use of silicon wafers for at least two reasons. First, a large wafer is essential to achieving the cost structure central to Moore's law. An additional consideration is the effective use of the tool base that will be in place when the new technology moves into advanced development.

The fabrication of III–V heterostructures on silicon has been under investigation for some time. The interest was fuelled by the integration of optical devices and CMOS logic circuits, multijunction solar cells, and the heterogeneous integration of CMOS circuitry and III–V electronic devices, among other applications. Nanometre-scale III–V CMOS transistors pose some unique requirements for heterogeneous integration. One is the need for a very thin buffer structure that converts the silicon lattice constant into the desired one. The thickness of the buffer layer matters for economic reasons, because long growth times limit process throughput, and for thermal reasons, because heat produced in the transistors must be effectively dissipated. Most buffer layers are made of ternary compound semiconductors, such as InAlAs or AlGaAs, which have poor thermal conductivity[86].

InGaAs MOSFETs with the lattice constant of InP have been fabricated on silicon by MBE using a 1.5-μm-thick composite GaAs/graded-InAlAs buffer layer[59]. Thinner buffers have been demonstrated for InGaAs HEMTs on silicon[87]. MOVPE, a more practical technique for manufacturing, is also being used. Excellent results have been obtained using an InP/InAlAs composite buffer around 1 μm thick[88].

A separate approach to the integration of III–V compounds on silicon is the transfer of a III–V device layer onto a silicon substrate that is covered by a thin dielectric[89,90]. This is similar to silicon-on-insulator (SOI), a well-established substrate in the silicon industry. The III–V-on-insulator approach even allows strain engineering of the transferred device layer[91]. The challenge for all transfer layer techniques is scaling up to large wafers.

Another silicon 'hetero-integration' technique is aspect ratio trapping (ART)[92]. This consists of the selective growth of lattice-mismatched material inside trenches with high aspect ratio and submicrometre dimensions. The trenches trap threading dislocations, yielding high-quality device layers. Ge, GaAs and InP films have been grown using ART. When combined with epitaxial lateral overgrowth (ELO), uniform high-quality films can be achieved[92]. GaAs MOSFETs with electrical characteristics comparable to those fabricated on GaAs substrates have been demonstrated using ART–ELO[93].

Although detailed studies of defect control have yet to be reported, it currently seems feasible to fabricate high-quality III–V-layer structures on silicon in a reasonably scalable manner. The greatest challenge is the preparation of a hybrid substrate for NMOS and PMOS transistors that incorporates islands of two different materials with different lattice constants, placed side by side with minimum overhead and yielding a planar surface. This is a critical problem that does not seem to be receiving sufficient attention.

Afterword

Moore's law is not a physical law in the manner of Gauss's law or Newton's laws of motion. It does not describe nature. Moore's law was formulated from observations of the exponential increase in transistor density in the early days of integrated electronics[94], but it has remained valid for 50 years. Moore's true insight was the understanding of the economics behind microelectronics[95]. The driver is not shrinking transistor size per se, but diminishing transistor cost. Transistor footprint scaling makes this possible, but only up to the point at which increased complexity starts eroding manufacturing yields. Moore's law is all about economics and human innovation, and silicon integrated electronics is a dramatic manifestation of the human spirit. But there is nothing unique about silicon. In the not too distant future it may no longer make economic sense to shrink silicon transistors further. It is then that III–V compounds could become the key for continuing Moore's law. Here I have reviewed some of the most pressing technical challenges that need to be overcome to make this happen and discussed some of the remarkable progress already made. ■

1. Ferain, I., Colinge, C. A. & Colinge, J.-P. Multigate transistors as the future of classical metal–oxide–semiconductor field-effect transistors. *Nature* **479,** 310–316 (2011).
2. Chau, R., Doyle, B., Datta, S., Kavalieros, J. & Zhang, K. Integrated nanoelectronics for the future. *Nature Mater.* **6,** 810–812 (2007).
3. Iwai, H. Roadmap for 22 nm and beyond. *Microelectron. Eng.* **86,** 1520–1528 (2009).

4. Frank, D. J. Power-constrained CMOS scaling limits. *IBM J. Res. Dev.* **46,** 235–244 (2002).
5. Theis, T. N. & Solomon, P. M. In quest of the "next switch": prospects for greatly reduced power dissipation in a successor to the silicon field-effect transistor. *Proc. IEEE* **98,** 2005–2014 (2010).
6. del Alamo, J. A. The high-electron mobility transistor at 30: impressive accomplishments and exciting prospects. *Int. Conf. Compound Semicond. Manuf. Technol.* 17–22 (CS ManTech, 2011).
7. ITRS, International Technology Working Groups *International Technology Roadmap for Semiconductors* ⟨http://www.itrs.net/Links/2010ITRS/Home2010.htm⟩ (ITRS, 2010).
8. Kim, D.-H. *et al.* 50-nm E-mode $In_{0.7}Ga_{0.3}As$ PHEMTs on 100-mm InP substrate with f_{max} > 1 THz. *IEEE Int. Electron Devices Meet.* 692–695 (IEEE, 2010).
9. Kim, D.-H. & del Alamo, J. A. 30-nm InAs PHEMTs With f_T = 644 GHz and f_{max} = 681 GHz. *IEEE Electron Device Lett.* **31,** 806–808 (2010).
10. Leuther, A. *et al.* 20 nm Metamorphic HEMT with 660 GHz f_T. *Int. Conf. Indium Phosphide Relat. Mater.* (IEEE, 2011).
11. Jeong, C., Antoniadis, D. A. & Lundstrom, M. S. On backscattering and mobility in nanoscale silicon MOSFETs. *IEEE Trans. Electron Devices* **56,** 2762–2769 (2009).
12. del Alamo, J. A., Kim, D.-H., Kim, T.-W., Jin, D. & Antoniadis, D. A. III–V CMOS: what have we learned from HEMTs? *Int. Conf. Indium Phosphide Relat. Mater.* (IEEE, 2011).
13. Kim, D. H., del Alamo, J. A., Antoniadis, D. A. & Brar, B. Extraction of virtual-source injection velocity in sub-100 nm III–V HFETs. *IEEE Int. Electron Devices Meet.* 861–864 (IEEE, 2009).
This paper reports that the electron injection velocity in InGaAs and InAs HEMTs is more than double that in Si MOSFETs of similar gate length at half the voltage.
14. Khakifirooz, A. & Antoniadis, D. A. MOSFET performance scaling — part I: historical trends. *IEEE Trans. Electron Devices* **55,** 1391–1400 (2008).
15. Liu, Y. *et al.* in *Fundamentals of III–V Semiconductor MOSFETs* (eds Oktyabrsky, S. & Ye, P. D.) 31–50 (Springer, 2010).
16. Fischetti, M. V. & Laux, S. E. Are GaAs MOSFETs worth building? A model-based comparison of Si and GaAs n-MOSFETs. *IEEE Int. Electron Devices Meet.* 481–484 (IEEE, 1989).
17. Jin, D., Kim, D.-H., Kim, T. & del Alamo, J. A. Quantum capacitance in scaled down III–V FETs. *IEEE Int. Electron Devices Meet.* 495–498 (IEEE, 2009).
18. Kim, T.-W., Kim, D.-H. & del Alamo, J. A. Logic characteristics of 40 nm thin-channel InAs HEMTs. *Int. Conf. Indium Phosphide Relat. Mater.* 496–499 (IEEE, 2010).
19. Kim, T.-W. & del Alamo, J. A. Injection velocity in thin-channel InAs HEMTs. *Int. Conf. Indium Phosphide Relat. Mater.* (IEEE, 2011).
20. Kim, D.-H. & del Alamo, J. A. 30 nm E-mode InAs PHEMTs for THz and future logic applications. *IEEE Int. Electron Devices Meet.* 719–722 (IEEE, 2008).
21. Dewey, G. *et al.* Logic performance evaluation and transport physics of Schottky-gate III–V compound semiconductor quantum well field effect transistors for power supply voltages (V_{CC}) ranging from 0.5 V to 1.0 V. *IEEE Int. Electron Devices Meet.* 487–490 (IEEE, 2009).
22. Umansky, V. *et al.* MBE growth of ultra-low disorder 2DEG with mobility exceeding 35 × $10^6 cm^2$/Vs. *J. Cryst. Growth* **311,** 1658–1661 (2009).
23. Spicer, W. E., Lindau, I., Skeath, P. & Su, C. Y. Unified defect model and beyond. *J. Vac. Sci. Technol.* **17,** 1019–1027 (1980).
24. Hinkle, C. L. *et al.* GaAs interfacial self-cleaning by atomic layer deposition. *Appl. Phys. Lett.* **92,** 071901 (2008).
25. Scarrozza, M. *et al.* A theoretical study of the initial oxidation of the GaAs(001)-β2(2×4) surface. *Appl. Phys. Lett.* **95,** 253504 (2009).
26. Becke, H., Hall, R. & White, J. Gallium arsenide MOS transistors. *Solid State Electron.* **8,** 813–818 (1965).
27. Mimura, T., Odani, K., Yokoyama, N., Nakayama, Y. & Fukuta, M. GaAs microwave MOSFETs. *IEEE Trans. Electron Devices* **25,** 573–579 (1978).
28. Passlack, M. *et al. In-situ* Ga_2O_3 process for GaAs inversion/accumulation device and surface passivation applications. *IEEE Int. Electron Devices Meet.* 383–386 (IEEE, 1995).
29. Passlack, M., Hong, M. & Mannaerts, J. P. Quasistatic and high frequency capacitance–voltage characterization of Ga_2O_3–GaAs structures fabricated by *in situ* molecular beam epitaxy. *Appl. Phys. Lett.* **68,** 1099–1101 (1996).
30. Passlack, M., Hong, M., Mannaerts, J. P., Kwo, J. R. & Tu, L. W. Recombination velocity at oxide–GaAs interfaces fabricated by *in situ* molecular beam epitaxy. *Appl. Phys. Lett.* **68,** 3605–3607 (IEEE, 1996).
31. Ren, F. *et al.* Enhancement-mode p-channel GaAs MOSFETs on semi-insulating substrates. *IEEE Int. Electron Devices Meet.* 943–945 (IEEE, 1996).
32. Ren, F. *et al.* Demonstration of enhancement-mode p- and n-channel GaAs MOSFETS with $Ga_2O_3(Gd_2O_3)$ As gate oxide. *Solid State Electron.* **41,** 1751–1753 (1997).
33. Ye, P. D. *et al.* GaAs MOSFET with oxide gate dielectric grown by atomic layer deposition. *IEEE Electron Device Lett.* **24,** 209–211 (2003).
34. Frank, M. M. *et al.* HfO_2 and Al_2O_3 gate dielectrics on GaAs grown by atomic layer deposition. *Appl. Phys. Lett.* **86,** 152904 (2005).
35. Huang, M. L. *et al.* Surface passivation of III–V compound semiconductors using atomic-layer-deposition-grown Al_2O_3. *Appl. Phys. Lett.* **87,** 252104 (2005).
36. Milojevic, M. *et al.* Half-cycle atomic layer deposition reaction studies of Al_2O_3 on $(NH_4)_2S$ passivated GaAs(100) surfaces. *Appl. Phys. Lett.* **93,** 252905 (2008).
37. Xuan, Y., Lin, H. C., Ye, P. D. & Wilk, G. D. Capacitance–voltage studies on enhancement-mode InGaAs metal-oxide-semiconductor field-effect transistor using atomic-layer-deposited Al_2O_3 gate dielectric. *Appl. Phys. Lett.* **88,** 263518 (2006).
38. Li, N. *et al.* Properties of InAs metal-oxide-semiconductor structures with atomic-layer-deposited Al_2O_3 dielectric. *Appl. Phys. Lett.* **92,** 143507 (2008).
39. Wu, Y. Q., Xuan, Y., Ye, P. D., Cheng, Z. & Lochtefeld, A. Inversion-type enhancement-mode InP MOSFETs with ALD Al_2O_3, HfO_2 and HfAlO nanolaminates as high-*κ* gate dielectrics. *IEEE Device Res. Conf.* 117–118 (IEEE, 2007).
40. Wang, W., Xiong, K., Wallace, R. M. & Cho, K. Impact of interfacial oxygen content on bonding, stability, band offsets, and interface states of GaAs:HfO_2 interfaces. *J. Phys. Chem. C* **114,** 22610–22618 (2010).
41. Lin, L. & Robertson, J. Defect states at III–V semiconductor oxide interfaces. *Appl. Phys. Lett.* **98,** 082903 (2011).
This paper provides density-functional theory calculations of interface defect states between high-*κ* oxides and GaAs, InAs and InP that are consistent with experimental observation.
42. Wang, W., Lee, G., Huang, M., Wallace, R. M. & Cho, K. First-principles study of GaAs(001)-β2(2×4) surface oxidation and passivation with H, Cl, S, F, and GaO. *J. Appl. Phys.* **107,** 103720 (2010).
43. Passlack, M., Droopad, R. & Brammertz, G. Suitability study of oxide/gallium arsenide interfaces for MOSFET applications. *IEEE Trans. Electron Devices* **57,** 2944–2956 (2010).
This experimental study of oxide–GaAs interfaces for MOSFET applications finds that whereas *in situ*-deposited Ga_2O_3 leads to excellent interfacial quality, *ex situ* ALD-deposited Al_2O_3 on GaAs leads to Fermi-level pinning.
44. Trinh, H. D. *et al.* The influences of surface treatment and gas annealing conditions on the inversion behaviors of the atomic-layer-deposition Al_2O_3/n-$In_{0.53}Ga_{0.47}As$ metal-oxide-semiconductor capacitor. *Appl. Phys. Lett.* **97,** 042903 (2010).
45. O'Connor, E. *et al.* A systematic study of $(NH_4)_2S$ passivation (22%, 10%, 5%, or 1%) on the interface properties of the $Al_2O_3/In_{0.53}Ga_{0.47}As$/InP system for n-type and p-type $In_{0.53}Ga_{0.47}As$ epitaxial layers. *J. Appl. Phys.* **109,** 024101 (2011).
46. Ok, I. & Lee, J. C. in *Fundamentals of III–V Semiconductor MOSFETs* (eds Oktyabrsky, S. & Ye, P. D.) 307–348 (Springer, 2010).
47. Wu, Y. D. *et al.* Engineering of threshold voltages in molecular beam epitaxy-grown $Al_2O_3/Ga_2O_3(Gd_2O_3)/In_{0.2}Ga_{0.8}As$. *J. Vac. Sci. Technol. B* **28,** C3H10–C3H13 (2010).
48. Cheng, C.-W., Apostolopoulos, G. & Fitzgerald, E. A. The effect of interface processing on the distribution of interfacial defect states and the C–V characteristics of III–V metal-oxide-semiconductor field effect transistors. *J. Appl. Phys.* **109,** 023714 (2011).
49. Chen, Y.-T. *et al.* Fluorinated HfO_2 gate dielectric engineering on $In_{0.53}Ga_{0.47}As$ metal-oxide-semiconductor field-effect-transistors. *Appl. Phys. Lett.* **96,** 103506 (2010).
50. Engel-Herbert, R., Hwang, Y., Cagnon, J. I. & Stemmer, S. Metal-oxide-semiconductor capacitors with ZrO_2 dielectrics grown on $In_{0.53}Ga_{0.47}As$ by chemical beam deposition. *Appl. Phys. Lett.* **95,** 062908 (2009).
51. Liu, Y., Xu, M., Heo, J., Ye, P. D. & Gordon, R. G. Heteroepitaxy of single-crystal $LaLuO_3$ on GaAs(111)A by atomic layer deposition. *Appl. Phys. Lett.* **97,** 162910 (2010).
52. Xu, M. *et al.* New insight into Fermi-level unpinning on GaAs: impact of different surface orientations. *IEEE Int. Electron Devices Meet.* 865–868 (IEEE, 2009).
53. Ishii, H. *et al.* High electron mobility metal–insulator–semiconductor field-effect transistors fabricated on (111)-oriented InGaAs channels. *Appl. Phys. Express* **2,** 121101 (2009).
54. Wang, C., Xu, M., Colby, R., Stach, E. A. & Ye, P. D. "Zero" drain-current drift of inversion-mode NMOSFET on InP (111)A surface. *IEEE Device Res. Conf.* 93–94 (IEEE, 2011).
55. Ye, P. D., Xuan, Y., Wu, Y. Q. & Xu, M. Inversion-mode $In_xGa_{1-x}As$ MOSFETs (x = 0.53, 0.65, 0.75) with atomic-layer-deposited high-*κ* dielectrics. *ECS Trans.* **19,** 605–614 (2009).
This paper reports significant improvements in ALD oxide InGaAs MOSFET transistor characteristics as the InAs mole fraction in the channel is increased.
56. Sonnet, A. M. *et al.* On the calculation of effective electric field in $In_{0.53}Ga_{0.47}As$ surface channel metal-oxide-semiconductor field-effect-transistors. *Appl. Phys. Lett.* **98,** 193501 (2011).
57. Sonnet, A. M. *et al.* Remote phonon and surface roughness limited universal electron mobility of $In_{0.53}Ga_{0.47}As$ surface channel MOSFETs. *Microelectron. Eng.* **88,** 1083–1086 (2011).
58. Oktyabrsky, S. *et al.* Electron scattering in buried InGaAs/high-*κ* MOS channels. *ECS Trans.* **35,** 385–395 (2011).
59. Radosavljevic, M. *et al.* Advanced high-*κ* gate dielectric for high-performance short-channel $In_{0.7}Ga_{0.3}As$ quantum well field effect transistors on silicon substrate for low power logic applications. *IEEE Int. Electron Devices Meet.* 319–322 (IEEE, 2009).
This paper reports the best logic performance of a III–V MOSFET so far, exceeding that of state-of-the-art silicon MOSFETs at 0.5 V.
60. Kuhn, K. J., Liu, M. Y. & Kennel, H. Technology options for 22 nm and beyond. *Int. Workshop Junct. Technol.* 1–6 (IEEE, 2010).
61. Waldron, N., Kim, D.-H. & del Alamo, J. A. A self-aligned InGaAs HEMT architecture for logic applications. *IEEE Trans. Electron Devices* **57,** 297–304 (2010).
62. Kim, T.-W., Kim, D.-H. & del Alamo, J. A. 60 nm Self-aligned-gate InGaAs HEMTs with record high-frequency characteristics. *IEEE Int. Electron Devices Meet.* 696–699 (IEEE, 2010).
63. Singisetti, U. *et al.* Ultralow resistance *in situ* ohmic contacts to InGaAs/InP. *Appl. Phys. Lett.* **93,** 183502 (2008).

This paper reports Mo/n+–InGaAs ohmic contacts with contact resistance comparable to state-of-the-art silicon technology.

64. Chin, H.-C., Gong, X., Liu, X. & Yeo, Y. Lattice-mismatched $In_{0.4}Ga_{0.6}As$ source/drain stressors with *in situ* doping for strained $In_{0.53}Ga_{0.47}As$ channel n-MOSFETs. *IEEE Electron Device Lett.* **30,** 805–807 (2009).
65. Terao, R. *et al.* InP/InGaAs composite metal–oxide–semiconductor field-effect transistors with regrown source and Al_2O_3 gate dielectric exhibiting maximum drain current exceeding 1.3 mA/μm. *Appl. Phys. Express* **4,** 054201 (2011).
This paper reports that a 100-nm gate length InGaAs MOSFET with raised epitaxially grown source and drain regions has excellent electrical characteristics.
66. Doyle, B. S. *et al.* High performance fully-depleted tri-gate CMOS transistors. *IEEE Electron Device Lett.* **24,** 263–265 (2003).
67. Wu, Y. Q., Xu, M., Wang, R. S., Koybasi, O. & Ye, P. D. High performance deep-submicron inversion-mode InGaAs MOSFETs with maximum G_m exceeding 1.1 mS/μm: new HBr pretreatment and channel engineering. *IEEE Int. Electron Devices Meet.* 323–326 (IEEE, 2009).
68. Radosavljevic, M. *et al.* Non-planar, multi-gate InGaAs quantum well field effect transistors with high-κ gate dielectric and ultra-scaled gate-to-drain/gate-to-source separation for low power logic applications. *IEEE Int. Electron Devices Meet.* 126–129 (IEEE, 2010).
69. Pearton, S. J. & Norton, D. P. Dry etching of electronic oxides, polymers, and semiconductors. *Plasma Process. Polym.* **2,** 16–37 (2005).
70. Hashemi, P., Gomez, L., Canonico, M. & Hoyt, J. L. Electron transport in gate-all-around uniaxial tensile strained-Si nanowire n-MOSFETs. *IEEE Int. Electron Devices Meet.* 1–14 (IEEE, 2008).
71. Suk, S. D. *et al.* High performance 5nm radius twin silicon nanowire MOSFET (TSNWFET): fabrication on bulk Si wafer, characteristics, and reliability. *IEEE Int. Electron Devices Meet.* 717–720 (IEEE, 2005).
72. Do, Q. T., Blekker, K., Regolin, I., Prost, W. & Tegude, F. J. Single n-InAs nanowire MIS-field-effect transistor: experimental and simulation results. *IEEE Int. Indium Phosphide Relat. Mater.* 392–395 (IEEE, 2007).
73. Egard, M. *et al.* Vertical InAs nanowire wrap gate transistors with $f_t > 7$ GHz and $f_{max} > 20$ GHz. *Nano Lett.* **10,** 809–812 (2010).
74. Hock, G., Hackbarth, T., Erben, U., Kohn, E. & Konig, U. High performance 0.25 μm p-type Ge/SiGe MODFETs. *Electron. Lett.* **34,** 1888–1889 (1998).
75. Bennett, B. R., Ancona, M. G., Boos, J. B. & Shanabrook, B. V. Mobility enhancement in strained p-InGaSb quantum wells. *Appl. Phys. Lett.* **91,** 042104 (2007).
76. Bennett, B., Ancona, M., Boos, J., Canedy, C. & Khan, S. Strained GaSb/AlAsSb quantum wells for p-channel field-effect transistors. *J. Cryst. Growth* **311,** 47–53 (2008).
77. Radosavljevic, M. *et al.* High-performance 40 nm gate length InSb p-channel compressively strained quantum well field effect transistors for low-power ($V_{CC} = 0.5$ V) logic applications. *IEEE Int. Electron Devices Meet.* 1–4 (IEEE, 2008).
78. Kuhn, K. J., Murthy, A., Kotlyar, R. & Kuhn, M. Past, present and future: SiGe and CMOS transistor scaling. *ECS Trans.* **33,** 3–17 (2010).
79. Xia, L., Boos, J. B., Bennett, B. R., Ancona, M. G. & del Alamo, J. A. Hole mobility enhancement in $In_{0.41}Ga_{0.59}Sb$ quantum-well field-effect transistors. *Appl. Phys. Lett.* **98,** 053505 (2011).
80. Xia, L. V. T., Oktyabrsky, S. & del Alamo, J. A. Mobility enhancement of two-dimensional hole Gas in an $In_{0.24}Ga_{0.76}As$ quantum well by <110> uniaxial strain. *Int. Symp. Compound Semicond. 2011* (IEEE, in the press).
81. Nainani, A. *et al.* Engineering of strained III–V heterostructures for high hole mobility. *IEEE Int. Electron Devices Meet.* 857–860 (IEEE, 2009).
82. Gomez, L., Chleirigh, C. N., Hashemi, P. & Hoyt, J. L. Enhanced hole mobility in high Ge content asymmetrically strained-SiGe p-MOSFETs. *IEEE Electron Device Lett.* **31,** 782–784 (2010).
83. Passlack, M. *et al.* Self-aligned GaAs p-channel enhancement mode MOS heterostructure field-effect transistor. *IEEE Electron Device Lett.* **23,** 508–510 (2002).
84. Nainani, A. *et al.* Development of high-κ dielectric for antimonides and a sub 350 °C III–V pMOSFET outperforming germanium. *IEEE Int. Electron Devices Meet.* 138–141 (IEEE, 2010).
This paper describes p-type InGaSb MOSFETs with an ALD Al_2O_3 gate dielectric and excellent characteristics.
85. Pillarisetty, R. *et al.* High mobility strained germanium quantum well field effect transistor as the p-channel device option for low power ($V_{cc} = 0.5$ V) III–V CMOS architecture. *IEEE Int. Electron Devices Meet.* 150–153 (IEEE, 2010).
86. Nakwaski, W. Thermal conductivity of binary, ternary, and quaternary III–V compounds. *J. Appl. Phys.* **64,** 159–166 (1988).
87. Hudait, M. K. *et al.* Heterogeneous integration of enhancement mode $In_{0.7}Ga_{0.3}As$ quantum well transistor on silicon substrate using thin (≤ 2 μm) composite buffer architecture for high-speed and low-voltage (0.5 V) logic applications. *IEEE Int. Electron Devices Meet.* 625–628 (2007).
88. Tang, C. W., Li, H., Zhong, Z., Ng, K. L. & Lau, K. M. Hetero-epitaxy of III–V compounds lattice-matched to InP by MOCVD for device applications. *IEEE Int. Conf. Indium Phosphide Relat. Mater.* 136–139 (IEEE, 2009).
89. Yokoyama, M. *et al.* Extremely-thin-body InGaAs-on-insulator MOSFETs on Si fabricated by direct wafer bonding. *IEEE Int. Electron Devices Meet.* 46–49 (2010).
90. Ko, H. *et al.* Ultrathin compound semiconductor on insulator layers for high-performance nanoscale transistors. *Nature* **468,** 286–289 (2010).
This paper reports on thin-channel InAs-on-insulator MOSFETs with excellent electrical characteristics fabricated on a silicon substrate by a layer-transfer process.
91. Fang, H. *et al.* Strain engineering of epitaxially transferred, ultrathin layers of III–V semiconductor on insulator. *Appl. Phys. Lett.* **98,** 012111 (2011).
92. Fiorenza, J. *et al*. Aspect ratio trapping: a unique technology for integrating Ge and III–Vs with silicon CMOS. *ECS Trans.* **33,** 963–976 (2010).
93. Wu, Y. Q. *et al.* Atomic-layer-deposited Al_2O_3/GaAs metal-oxide-semiconductor field-effect transistor on Si substrate using aspect ratio trapping technique. *Appl. Phys. Lett.* **93,** 242106 (2008).
94. Moore, G. E. Cramming more components onto integrated circuits. *Electronics* **38,** 114–117 (1965).
95. Hutcheson, G. D. in *Into the Nano Era: Moore's Law Beyond Planar Silicon CMOS* (ed. Huff, H.) 11–38 (Springer, 2009).

Acknowledgements I have enjoyed stimulating discussions with students, collaborators and colleagues. I am particularly thankful to D. Antoniadis, R. Chau, S. Datta, J. Hoyt, D. Jin, D.-H. Kim, T.-W. Kim, A. Kummel, J. Lin, M. Lundstrom, S. Oktyabrsky, M. Passlack, M. Radosavljevic, E. Vogel, N. Waldron, R. Wallace, L. Xia and P. Ye. Research on III–V CMOS transistors at my lab at MIT has been funded by the Materials, Structures and Devices FCRP Center and Intel Corporation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to the author (alamo@mit.edu).

doi:10.1038/nature10678

Academic and industry research progress in germanium nanodevices

Ravi Pillarisetty[1]

Silicon has enabled the rise of the semiconductor electronics industry, but it was not the first material used in such devices. During the 1950s, just after the birth of the transistor, solid-state devices were almost exclusively manufactured from germanium. Today, one of the key ways to improve transistor performance is to increase charge-carrier mobility within the device channel. Motivated by this, the solid-state device research community is returning to investigating the high-mobility material germanium. Germanium-based transistors have the potential to operate at high speeds with low power requirements and might therefore be used in non-silicon-based semiconductor technology in the future.

Gordon Moore's visionary prediction, made in 1965, that the number of transistors on an integrated circuit chip would double every two years continues to be the main idea guiding the semiconductor industry. With each new generation of technology, not only does the manufacturing process become more economical, but the individual transistors also become smaller and faster, and require less power. Whereas the first microprocessor consisted of 2,300 transistors, today's state-of-the-art microprocessors are made up of more than 1 billion transistors, which operate 50,000 times faster than the early transistors[1].

In the past decade, the semiconductor industry has implemented several innovations to help continue to meet Moore's law. For example, in 2003, the 90-nm technology node used strain engineering to further increase the performance of silicon transistors[2]. The application of strain to the silicon channel considerably improved charge-carrier mobility, which translated directly into an increase in drive current. Subsequently, the fundamental issue of dielectric scaling in silicon dioxide was addressed, in 2007, by introducing a novel high scaling factor (high-κ) metal gate stack into the 45-nm node technology[3]. These silicon transistors with high-κ metal gates had a significantly lower electrical gate oxide thickness (TOXE), as well as better scalability and performance, than the previous, 65-nm node, SiO_2-based technology.

Recently, there has been considerable interest in the solid-state device community in researching the use of non-silicon materials that have high-mobility charge carriers, to replace the current silicon-based transistor channel[4–43]. Increasing the carrier mobility can improve the transistor drive current, which can be used to improve device performance or to maintain performance and reduce power consumption. Germanium has the highest p-type mobility of all of the known semiconductor materials and is therefore an attractive option as a silicon replacement in future low-power logic applications. Germanium is not new to the semiconductor industry, however. The transistor and the integrated circuit, which were developed in 1947 and 1958, respectively, are the foundations of today's US$260 billion semiconductor industry, and it is often overlooked that the initial groundbreaking studies were conducted in germanium rather than in silicon, which dominates the semiconductor market. In fact, the initial years of the solid-state industry involved germanium diode and bipolar-junction-transistor technology almost exclusively. It was not until the mid-1960s, after the discovery of SiO_2 dielectric passivation[44] and the planar metal–oxide–semiconductor field-effect transistor (MOSFET) process[45], that silicon took over as the dominant industry material.

In the past decade, interest in germanium devices has undergone a resurgence as the research community has re-investigated this material as a possible high-mobility replacement for the mainstream silicon MOSFET technology[13–43]. In this Review, I examine the history and recent progress of industry and academic research into the use of germanium channel materials as a replacement for silicon-based p-type MOSFETs (PMOSs). Such research could lead to a non-silicon transistor architecture, based on low-power group III–V/Ge complementary metal–oxide–semiconductors (CMOSs).

High-mobility non-silicon channel materials

Figure 1 summarizes the semiconductor materials landscape by showing the relationship between the bulk electron and hole mobility and the bandgap for silicon, germanium and a variety of group III–V compound semiconductor materials. The bandgap is an important parameter for selecting a transistor material because it affects both the supply voltage that the device can operate at and the scalability of the device. For example, if the bandgap is too large, the lack of metal with a suitable work function prevents the threshold voltage of the device from being small enough to allow operation at low supply voltage. By contrast, if the bandgap is too small, the behaviour of the device will become degraded because of off-state leakage currents resulting from both conventional thermionic emission and band-to-band tunnelling. Hence, having either too large or too small a bandgap will affect the scalability of the transistor.

Figure 1 shows that the electron (n-type) mobility in III–V materials provides a considerable advantage over that of silicon, with several materials showing an n-type mobility of more than 10,000 $cm^2 V^{-1} s^{-1}$. Over the past few years, considerable research progress has been made in demonstrating the viability of III–V channel materials as a replacement for silicon in low-power logic applications[4–7]. Several key research breakthroughs have been made: these include the integration of such materials on a silicon substrate[5], and the development of a high-quality gate-dielectric interface with a scaled effective oxide thickness, a low interface state density (D_{it}) and controlled short-channel effects[6]. Furthermore, InGaAs transistors have been shown to have considerable drive current gains over state-of-the-art silicon at low supply voltages, owing to their high carrier mobility[7]. These results bolster the proposition that the substantial mobility gains obtained by using III–V channel materials can provide major performance and power-consumption improvements over silicon devices for future technology nodes.

One of the key challenges faced by researchers studying alternative

[1]Components Research, Technology and Manufacturing Group, Intel Corporation, 5200 Northeast Elam Young Parkway, Hillsboro, Oregon 97124, USA.

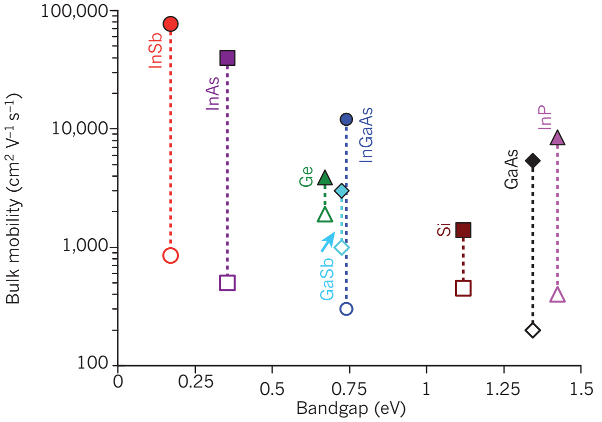


Figure 1 | The mobility landscape of semiconductors. The bulk mobility is plotted against the bandgap for silicon, germanium and a variety of group III–V materials. Filled symbols indicate electrons, and open symbols indicate holes. Germanium offers the highest hole mobility of any known semiconductor material.

channel materials is identifying a PMOS solution, which could boost device performance to a significantly higher level than that of the current state-of-the-art strained silicon PMOS transistors. Although the electron mobility in several III–V materials can be very high, with n-type MOSFETs (NMOSs) made of those materials significantly outperforming silicon NMOS devices, the corresponding hole mobility is significantly lower than the electron mobility. This difference, which is shown for a variety of III–V materials in Fig. 1, arises from their considerably larger valence-band effective mass, together with their higher carrier-scattering rate for hole transport than for electrons. Given the current CMOS-based design for logic technology, which uses both NMOS and PMOS transistors, researchers need to find ways to markedly improve non-silicon PMOS device performance so that microprocessor performance gains can be maximized.

On the basis of the data in Fig. 1, the most promising material for the PMOS device in a future III–V-based CMOS architecture is not a III–V-based material. Of all of the known semiconductor materials, germanium has the highest hole mobility, which is about twofold higher than the best III–V-based p-type materials. Furthermore, non-silicon CMOS technology would be targeted for use at low supply voltages, and at the low supply voltage of about 0.5 V, the leakage from band-to-band tunnelling and thermionic emission should not have a significant impact on germanium-based devices, owing to germanium's large bandgap of 0.67 eV relative to the supply voltage. To address the viability of germanium as a future channel material, researchers have investigated several key areas — the heterogeneous integration of germanium on a silicon substrate, the use of conventional MOSFET and quantum-well device architectures, a suitable gate-dielectric passivation scheme on germanium, and the carrier transport and short-channel performance of germanium — each of which is discussed below.

Heterogeneous integration of germanium on silicon

For a germanium-based technology to be compatible with today's conventional CMOS manufacturing process, the integration of germanium on a silicon substrate is imperative. By incorporating the germanium transistor channel on a silicon wafer, this new material can be plugged into the mainstream silicon CMOS platform, eliminating the considerable increase in manufacturing costs that would be associated with developing new types of non-silicon substrates. However, the materials growth challenges in developing such a consolidation technique are substantial because the lattice constant of germanium is 4% larger than that of silicon. Any germanium growth that is initiated on silicon will start to relax almost immediately through the formation of misfit dislocations, because the critical thickness for such a structure is only a few monolayers. The extremely high number of defects formed in such a thin germanium layer would considerably degrade carrier mobility and augment junction leakage, making transistor operation impossible.

To accommodate the large lattice mismatch between the germanium and silicon materials, a thick Si_xGe_{1-x} buffer layer can be grown between the silicon substrate and the active germanium layer, whose defect density is significantly reduced by such a process[38–42,46]. This buffered-growth technique provides a way of building germanium devices on a silicon substrate and benchmarking their electrical properties (such as mobility, the ratio of the on current to the off-state leakage current, and carrier velocity) against those of conventional silicon MOSFETs. The buffers are usually several micrometres thick and comprise a Si_xGe_{1-x} material whose composition is either gradually graded or abruptly changed in a few steps to accommodate the lattice mismatch. It is also possible to grow germanium directly onto silicon, although the germanium must be grown in a very thick layer so that it has a reasonably low defect density at the surface. A variety of growth techniques, such as molecular beam epitaxy[38–41], ultra-high-vacuum chemical vapour deposition[42,46] and chemical vapour deposition[19,21,22,24], have been investigated for growing germanium on silicon substrates.

Although buffered-growth techniques for integrating germanium on silicon have made considerable progress in recent years, much additional research and development is needed before a manufacturable germanium-on-silicon scheme is possible. At present, the best SiGe buffer systems are several micrometres thick; therefore, they take a long time to grow, which increases process costs. To address this issue, several alternatives have been investigated. The aspect-ratio-trapping process has recently been studied for growing germanium and SiGe buffer layers within narrow oxide trenches patterned on silicon substrates[47–49]. This technique attempts to isolate threading dislocations along the oxide sidewalls, a process that could lead to a major reduction in buffer thickness. For example, recent work[48] (Fig. 2) shows that when growing germanium directly on silicon in 200-nm-wide trenches, a high-quality germanium surface can be achieved when the germanium thickness reaches 250 nm, which is significantly thinner than the buffer sizes that have been reported for blanket germanium-on-silicon growth[38–42,46]. In addition, several alternative techniques, including wafer bonding[50], condensation[51] and lateral liquid-phase epitaxy[52], have also been investigated for germanium on silicon integration.

MOSFET and quantum-well device architectures

Device research has mainly focused on two different architectures that can be built on these virtual germanium substrates: conventional MOSFETs[13–15,21,22,25,26,28,29,33,37], and quantum-well field-effect transistors (QWFETs)[16–20,23,24,27,38–42]. By growing a thick germanium layer directly on the silicon substrate or the Si_xGe_{1-x} buffer, the relaxed germanium surface on top can be used to process and fabricate a conventional MOSFET (Fig. 3a). These devices use n-type body doping and operate by forming a p-type inversion layer in the same manner as mainstream silicon MOSFETs. By contrast, a large amount of device research has focused on the QWFET (Fig. 3b). In this case, a thin germanium layer, which functions as the device channel, is grown biaxially strained to the Si_xGe_{1-x} buffer layer. The germanium layer is capped with a top barrier of either Si_xGe_{1-x} or silicon, such that the valence-band offsets on either side are able to quantum confine hole carriers to the germanium. In addition, modulation doping techniques can be used to supply holes to

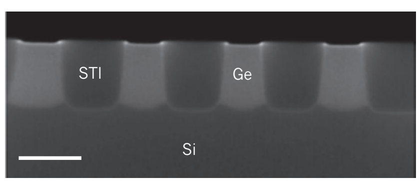


Figure 2 | The integration of germanium on silicon. A cross-sectional scanning electron microscopy image of germanium-on-silicon growth, using the aspect-ratio-trapping process[48]. Scale bar, 0.25 μm. Reproduced, with permission, from ref. 48. STI, shallow-trench isolation.

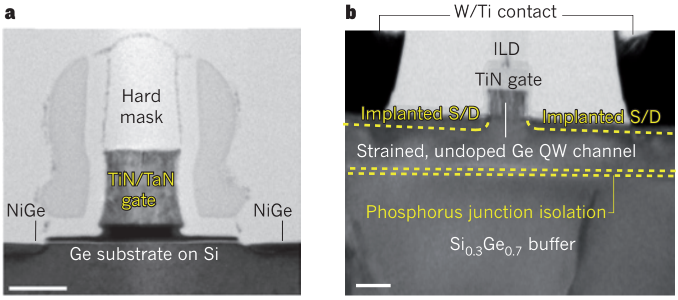


Figure 3 | The two main germanium device architectures. Germanium device research has primarily focused on two device architectures: the germanium MOSFET[21] (**a**); and the germanium QWFET[24] (**b**). In the germanium MOSFET shown (**a**), current flows between the two NiGe contacts. The flow of current is controlled by applying voltage to the TiN/TaN gate electrode. The hard mask used to pattern the gate electrode is indicated. In the germanium QWFET shown (**b**), current flows between the two W/Ti contacts, which contact the implanted source–drain (S/D) regions of the device. Current flow between the two source–drain regions travels through the high-mobility, strained, undoped germanium quantum-well (QW) channel and is modulated by applying voltage to the TiN gate electrode. The phosphorus-junction isolation layer suppresses parasitic off-state leakage current in the $Si_{0.3}Ge_{0.7}$ buffer. The interlayer dielectric (ILD) is also indicated. Scale bars, 50 nm (**a**), 100 nm (**b**). Panel **a** is reproduced, with permission, from ref. 21. Panel **b** is reproduced, with permission, from ref. 24.

the quantum-well channel, which remains undoped, because any ionized dopants would be placed in the Si_xGe_{1-x} barrier.

Unlike germanium MOSFET structures, which require full transistor processing for electrical characterization of the inversion layer's hole mobility, the mobility of the two-dimensional hole system in the germanium quantum-well structure can easily be determined immediately after growth by using Hall measurements. Such Hall measurements provide a direct measurement of the intrinsic two-dimensional channel mobility in a given material. Quantum-well structures that have been characterized by Hall measurements generally have fairly large top barrier layers, allowing the surface states and interface traps to be spaced far from the channel, thereby eliminating mobility degradation from these sources. By contrast, a transistor built using such a quantum-well structure must have an ultra-scaled gate dielectric that is very close to the channel for it to be as scalable as current silicon transistors. Mobility degradation would be observed in the transistor channel unless the dielectric interface was ideally engineered, with no impact from scattering related to the interface trap states. In that sense, the Hall mobility is a vital piece of information because it represents the highest possible mobility that could be obtained in a scaled transistor made from such a quantum-well structure. Benchmarking the Hall mobility of a variety of germanium and III–V p-type quantum-well structures[8–12,24,39–41] (Fig. 4a) shows that germanium has the highest hole mobility of all of the two-dimensional hole systems.

The gate dielectric

The discovery of a stable SiO_2 dielectric-passivation scheme on silicon[44] with a very low interface trap density by Atalla's research group at Bell Laboratories in 1959 was perhaps the most important finding in semiconductor device physics since the original invention of the transistor. Subsequently, this high-quality gate-dielectric process was used to fabricate the first field-effect transistor (FET)[45]. The silicon MOSFET was much more scalable than the bipolar junction transistor and went on to become the key building block in integrated circuit and microprocessor manufacturing. Unlike silicon, however, germanium's native oxide is unstable and readily decomposes into several Ge_yO_x suboxides, which have a high density of dangling bonds at the interface. These interface trap states, which cannot be hydrogen passivated by using conventional forming gas anneals, can markedly degrade MOSFET performance[37,43]. The intrinsically high carrier mobility of germanium becomes significantly degraded because of carrier scattering arising from these traps. In addition, these interface trap states will also degrade the off-state leakage current and subthreshold turn-off of a germanium-based device, significantly affecting device scalability.

The challenge of finding a high-quality gate dielectric on germanium, with an interface trap density matched to that of state-of-the-art silicon, has been a fundamental problem in the semiconductor industry for more than 50 years. Unlike the high-quality Si/SiO_2 interface, which could be thermally grown and passivated using annealing techniques, germanium gate-dielectric research has faced much greater challenges. Only by finding such a gate-dielectric interface can the performance gains from germanium's high mobility be realized. The earliest studies of germanium MOSFET devices investigated chemical-vapour-deposition-based deposition of SiO_2 on germanium[13,14]. These initial studies, which

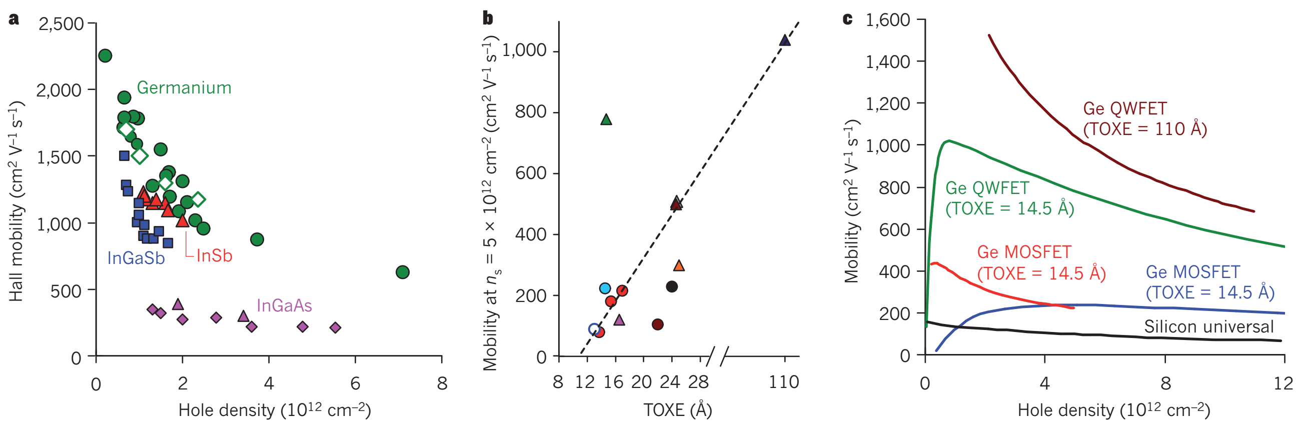


Figure 4 | Benchmarking germanium device mobility. **a**, The relationship between Hall mobility and hole density for biaxially strained quantum-well systems based on InSb (red triangles[8]), InGaAs (pink diamonds[9,11] and pink triangles[10]), InGaSb (blue squares[12]) and germanium (green circles[24] and green diamonds[38–40]). Germanium has the highest two-dimensional hole mobility of any known quantum-well system. **b**, The relationship between mobility at a carrier density (n_s) of 5×10^{12} cm^{-2} and TOXE for germanium devices in the literature[17–19,22–28,33]. These data highlight the problem of mobility degradation as the TOXE is scaled down. Quantum-well devices are indicated by triangles, and MOSFET devices are indicated by circles. From right to left and top to bottom, data are taken from the following: ref. 23 (blue triangle), ref. 24 (green triangle), ref. 17 (brown triangle), ref. 19 (purple triangle; almost superimposed with brown triangle), ref. 18 (orange triangle), ref. 25 (black circle), ref. 22 (red circles; three data points), ref. 33 (filled blue circle), ref. 26 (brown circle), ref. 27 (pink triangle) and ref. 28 (open blue circle). **c**, The relationship between mobility and hole density for state-of-the-art long-channel germanium MOSFET devices (blue curve[24] and red curve[33]) and QWFET devices (green curve[24] and brown curve[23]). The silicon universal mobility is provided as a reference point. The larger mobility observed in the QWFET devices arises from reduced impurity scattering and biaxial strain.

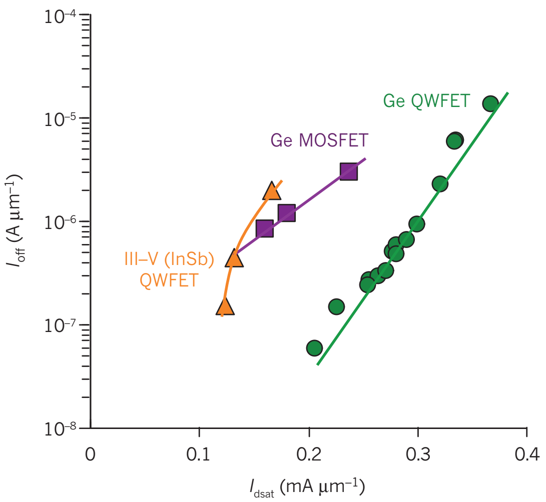


Figure 5 | Short-channel performance of p-type non-silicon devices. The relationship between the off-state leakage current (I_{off}) and the saturation drive current (I_{dsat}) at a supply voltage of 0.5 V for the state-of-the-art short-channel germanium QWFET[24], germanium MOSFET[21] and III–V (InSb) QWFET[8]. The germanium QWFET has the highest p-channel short-channel performance of any non-silicon device architecture. Reproduced, with permission, from ref. 24.

attempted to deposit a low-temperature SiO_2 dielectric on germanium, showed marked signatures of interface trap states, which significantly degraded carrier mobility. Subsequent work focused on thermally grown GeO_2 or GeON[15], which improved the interface trap density over that achieved using the chemical-vapour-deposition process but still did not approach the quality of the mainstream Si/SiO_2 interface. As an alternative to these approaches in which a dielectric interface was formed directly on germanium, the use of an epitaxial silicon capping layer, to move the dielectric interface outside the germanium, has also been investigated[16–24]. The main advantage of this technique is that conventional silicon dielectric techniques could be used with greater success, as the gate oxide is no longer being formed directly on germanium.

Over the past few years, there has been a marked renewal in germanium dielectric research in the semiconductor device community, with noteworthy progress being made on several fronts. Recent work on direct thermal growth of GeON using ozone-oxidation and nitridation techniques has achieved a low interface trap density and minimal degradation of carrier mobility in germanium MOSFETs[29]. Because the temperature and O_2 pressure are carefully controlled during growth, as are subsequent nitridation anneals, GeO_2-based dielectrics have recently been shown to have a low interface trap density[30,31]. Alternatively, transistors formed by using germanium condensation techniques, in which the gate dielectric is SiO_2 that is created during condensation, have been shown to have a reasonable channel mobility[32]. Furthermore, interface and transistor characterization of germanium devices with silicon-cap-based gate dielectrics has also demonstrated good capacitance–voltage characteristics with a low interface trap density[17–24]. In the gate stacks of these devices, the two-dimensional hole carriers are confined to the Ge/Si interface owing to the roughly 400-meV valence-band offset between these materials. Additionally, high-κ gate-dielectric materials such as HfO_2 and ZrO_2 have been successfully incorporated into such germanium gate stacks without degradation of the interface properties[19,21,22,24,25]. Researchers have also demonstrated promising capacitance–voltage characteristics using bilayer high-κ gate stacks such as $LaON/HfO_2$ (ref. 34) and Al_2O_3/TiO_2 (ref. 35). These recent studies clearly demonstrate an important advance in the field, establishing that high-quality dielectric interfaces can be engineered on germanium. However, before such gate stacks on germanium can be put to practical use in a manufacturable silicon-replacement technology, the thickness of the dielectric, which strongly affects transistor scalability and performance, remains a fundamental problem that needs to be resolved.

State-of-the-art silicon transistors have gate stacks with a high-κ dielectric and a TOXE of less than 14 Å[2]. There are two benefits of the scaled dielectric: the electrostatics are better, which improves scalability; and the gate capacitance is higher, which leads to more charge and drive current. A historical benchmarking of the data reported for the best germanium devices is provided in Fig. 4b. Although good dielectric interfaces with high mobility can be achieved on germanium, almost all of these devices have a TOXE significantly larger than today's mainstream silicon technology. Devices built with such gate stacks would have significantly degraded scalability and on-state charge density, eliminating any performance gains expected from the improvement in mobility over silicon. As the TOXE is scaled in these germanium gate stacks, the mobility decreases significantly[24,36]. This finding highlights the fundamental challenge that must be met if germanium is to become a viable technology. Germanium dielectric research must focus on finding new ways to scale the gate stack while maintaining high carrier mobility. In the past year, there have been considerable advances on this front in studies using silicon-cap[24] and GeO_x (ref. 33) transition layers between germanium and the high-κ dielectric.

Carrier transport and short-channel performance

A comparison of the long-channel transport in germanium devices is presented in Fig. 4c, showing the mobility and carrier density for state-of-the-art long-channel MOSFET[24,33] and QWFET devices[23,24]. All of these devices show significant mobility gains relative to the silicon universal mobility. At a scaled TOXE of 14.5 Å, the germanium QWFET has a mobility that is about fourfold higher than the germanium MOSFET. These mobility gains arise from both the reduction in impurity scattering (owing to the undoped quantum-well channel) and the enhancement in mobility (owing to biaxial strain). At a larger TOXE, 110 Å, even higher hole mobilities, of more than 1,000 $cm^2 V^{-1} s^{-1}$ at a carrier density (n_s) of $5 \times 10^{12} cm^{-2}$, have been achieved in a germanium QWFET[23]. In this study, the additional mobility enhancements were obtained using a lower germanium concentration SiGe buffer to impart a higher biaxial strain in the germanium QWFET.

Although research progress has been made in scaling the dielectric thickness on germanium and maintaining a significantly higher mobility than in silicon, there are several other key elements that are required for high-performance, short-channel germanium devices. These include source–drain implantation and contact schemes to mitigate parasitic contact resistance while maintaining short-channel effects. Traditionally, parallel conduction in the SiGe buffer has severely degraded the off-state leakage current in germanium QWFETs, although progress has recently been made towards eliminating its contribution to parasitic leakage[24]. Data for state-of-the-art, short-channel germanium devices are presented in Fig. 5, which shows the relationship between the saturation drive current (I_{dsat}) and the off-state leakage current (I_{off}) benchmarked at a supply voltage of 0.5 V. These data show that the germanium QWFET[24] has roughly twice the saturation drive current of the germanium MOSFET[21] and the III–V QWFET.

Perspectives

The semiconductor industry is continuing to improve transistor performance, but there has recently been a growing emphasis on also reducing transistor power consumption. To that end, there has been an increase in research on high-mobility materials for use as the transistor channel. A non-silicon-based CMOS architecture based on high-mobility materials could allow a significant reduction in power consumption while maintaining performance. Although n-channel transistors based on high-mobility III–V materials have shown significant performance gains over state-of-the-art silicon, the corresponding p-type transistor has not been demonstrated. However, there has been a recent resurgence in research on germanium, which has the highest hole mobility of any known semiconductor material. Considerable research progress has been made by industry and academia in several key areas such as integrating germanium on silicon, generating a high-quality gate dielectric on

germanium, and investigating germanium MOSFET and QWFET devices and their long- and short-channel transport. These recent results are encouraging, and the semiconductor device community remains excited that germanium could be a viable p-channel option for a future non-silicon CMOS architecture. ■

1. Moore, G. E. No exponential is forever: but 'forever' can be delayed! *Digest Tech. Papers Int. Solid-State Circuits Conf.* 20–23 (IEEE, 2003).
2. Ghani, T. *et al.* A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors. *Tech. Digest IEEE Electron Devices Meet.* 11.6.1–11.6.3 (IEEE, 2003).
3. Mistry, K. *et al.* A 45nm logic technology with high-k$^+$ metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging. *Tech. Digest IEEE Electron Devices Meet.* 247–250 (IEEE, 2007).
4. Kim, D. & Del Alamo, J. Scaling behavior of $In_{0.7}Ga_{0.3}As$ HEMTs for logic. *Tech. Digest IEEE Electron Devices Meet.* 837–841 (IEEE, 2006).
5. Hudait, M. *et al.* Heterogeneous integration of enhancement mode $In_{0.7}Ga_{0.3}As$ quantum well transistor on silicon substrate using thin (< 2 μm) composite buffer architecture for high-speed and low-voltage (0.5 V) logic applications. *Tech. Digest IEEE Electron Devices Meet.* 625–628 (IEEE, 2007).
6. Radosavljevic, M. *et al.* Advanced high-*κ* gate dielectric for high-performance short-channel $In_{0.7}Ga_{0.3}As$ quantum well field effect transistors on silicon substrate for low power logic applications. *Tech. Digest IEEE Electron Devices Meet.* 319–322 (IEEE, 2009).
7. Dewey, G. *et al.* Logic performance evaluation and transport physics of Schottky-gate III–V compound semiconductor quantum well field effect transistors for power supply voltages (V_{CC}) ranging from 0.5 V to 1.0 V. *Tech. Digest IEEE Electron Devices Meet.* 487–490 (IEEE, 2009).
8. Radosavljevic, M. *et al.* High-performance 40 nm gate length InSb p-channel compressively strained quantum well field effect transistors for low-power (V_{CC} = 0.5 V) logic applications. *Tech. Digest IEEE Electron Devices Meet.* 727–730 (IEEE, 2008).
9. Kudo, M., Matsumoto, H., Tanimoto, T., Mishima, T. & Ohbu, I. Improved hole transport properties of highly strained $In_{0.35}Ga_{0.65}As$ channel double-modulation-doped structures grown by MBE on GaAs. *J. Cryst. Growth* **175,** 910–914 (1997).
10. Nagaiah, P., Tokranov, V. & Oktyabrsky, S. Strained quantum wells for p-channel InGaAs CMOS. *Mater. Res. Soc. Symp. Proc.* 1108-A12-01 (Cambridge Univ. Press, 2009).
11. Schirber, J. E., Fritz, I. J. & Dawson, L. R. Light hole conduction in InGaAs/GaAs strained-layer superlattices. *Appl. Phys. Lett.* **46,** 187–189 (1985).
12. Bennett, B. R., Ancona, M. G., Boos, J. B. & Shanabrook, B. V. Mobility enhancement in strained p-InGaSb quantum wells. *Appl. Phys. Lett.* **91,** 042104 (2007).
13. Chang, L. L. & Yu, H. N. The germanium insulated-gate-field-effect transistor (FET). *Proc. IEEE* **5,** 316–317 (IEEE, 1965).
14. Wang, K. L. & Gray, P. V. Fabrication and characterization of germanium ion-implanted IGFET's. *IEEE Trans. Electron Devices* **22,** 353–355 (1975).
15. Martin, S. C., Hitt, L. M. & Rosenberg, J. J. p-Channel germanium MOSFET's with high channel mobility. *IEEE Electron Device Lett.* **10,** 325–326 (1989).
16. Lee, M. L. *et al.* Strained Ge channel p-type metal–oxide–semiconductor field-effect transistors grown on $Si_{1-x}Ge_x$/Si virtual substrates. *Appl. Phys. Lett.* **79,** 3344–3346 (2001).
17. Weber, O. *et al.* Strained Si and Ge MOSFETs with high-*k*/metal gate stack for high mobility dual channel CMOS. *Tech. Digest IEEE Electron Devices Meet.* 137–140 (IEEE, 2005).
18. Krishnamohan, T., Krivokapic, Z., Uchida, K., Nishi, Y. & Saraswat, K. C. Low defect ultra-thin fully strained-Ge MOSFET on relaxed Si with high mobility and low band-to-band-tunneling (BTBT). *Tech. Digest Papers Symp. VLSI Technol.* 82–83 (IEEE, 2005).
19. Nicholas, G. *et al.* High mobility strained Ge pMOSFETs with high-*κ*/metal gate. *IEEE Electron Device Lett.* **28,** 825–827 (2007).
20. Ni Chleirigh, C. *Strained SiGe-channel p-MOSFETs: Impact of Heterostructure Design and Process Technology*. PhD thesis, Mass. Inst. Technol. (2007).
21. Mitard, J. *et al.* Record I_{ON}/I_{OFF} performance for 65 nm Ge pMOSFET and novel Si passivation scheme for improved EOT scalability. *Tech. Digest IEEE Electron Devices Meet.* 873–876 (IEEE, 2008).
This article reports on state-of-the-art short-channel germanium MOSFET performance.
22. Mitard, J. *et al.* Impact of EOT scaling down to 0.85 nm on 70 nm Ge-pFETs technology with STI. *Tech. Digest Papers Symp. VLSI Technol.* 82–83 (IEEE, 2009).
23. Gomez, L., Ni Chleirigh, C., Hashemi, P. & Hoyt, J. L. Enhanced hole mobility in high Ge content asymmetrically strained-SiGe p-MOSFETs. *IEEE Electron Device Lett.* **31,** 782–784 (2010).
This article reports the highest mobility achieved so far in a germanium QWFET.
24. Pillarisetty, R. *et al.* High mobility strained germanium quantum well field effect transistor as the p-channel device option for low power (V_{CC} = 0.5 V) III–V CMOS architecture. *Tech. Digest IEEE Electron Devices Meet.* 150–153 (IEEE, 2010).
This article describes state-of-the-art mobility and short-channel performance in a germanium QWFET with a scaled TOXE.
25. Chui, C. *et al.* A sub-400 °C germanium MOSFET technology with high-*κ* dielectric and metal gate. *Tech. Digest IEEE Electron Devices Meet.* 437–440 (IEEE, 2002).
26. Huang, C. H. *et al.* Very low defects and high performance Ge-on-insulator p-MOSFETs with Al_2O_3 gate dielectrics. *Tech. Digest Papers Symp. VLSI Technol.* 119–120 (IEEE, 2003).
27. Ritenour, A. *et al.* Epitaxial strained germanium p-MOSFETs with HfO_2 gate dielectric and TaN gate electrode. *Tech. Digest IEEE Electron Devices Meet.* 433–436 (IEEE, 2003).
28. Whang, S. J. *et al.* Germanium p- & n-MOSFETs fabricated with novel surface passivation (plasma-PH_3 and thin AlN) and TaN/HfO_2 gate stack. *Tech. Digest IEEE Electron Devices Meet.* 307–310 (IEEE, 2004).
29. Kuzum, D. *et al.* Interface-engineered Ge (100) and (111), N- and P-FETs with high mobility. *Tech. Digest IEEE Electron Devices Meet.* 723–726 (IEEE, 2007).
30. Kita, K. *et al.* Comprehensive study of GeO_2 oxidation, GeO desorption and GeO_2–metal interaction: understanding of Ge processing kinetics for perfect interface control. *Tech. Digest IEEE Electron Devices Meet.* 693–696 (IEEE, 2009).
31. Wang, S. K. *et al.* Desorption kinetics of GeO from GeO_2/Ge structure. *J. Appl. Phys.* **108,** 054104 (2010).
32. Tezuka, T. *et al.* A new strained-SOI/GOI dual CMOS technology based on local condensation technique. *Tech. Digest Papers Symp. VLSI Technol.* 80–81 (IEEE, 2005).
33. Zhang, R., Iwasaki, T., Taoka, N., Takenaka, M. & Takagi, S. High mobility Ge pMOSFETs with ~ 1 nm thin EOT using Al_2O_3/GeO_x/Ge gate stacks fabricated by plasma post oxidation. *Tech. Digest Papers Symp. VLSI Technol.* 56–57 (IEEE, 2011).
This article describes state-of-the-art mobility in a germanium MOSFET with a scaled TOXE.
34. Xu, H. X., Xu, J. P., Li, C. X. & Lai, P. T. Improved electrical properties of Ge metal–oxide–semiconductor capacitors with high-*κ* HfO_2 gate dielectric by using La_2O_3 interlayer sputtered with/without N_2 ambient. *Appl. Phys. Lett.* **97,** 022903 (2010).
35. Swaminathan, S., Shandalov, M., Oshima, Y. & McIntyre, P. C. Bilayer metal oxide gate insulators for scaled Ge-channel metal–oxide–semiconductor devices. *Appl. Phys. Lett.* **96,** 082904 (2010).
36. Caymax, M. *et al.* Germanium for advanced CMOS anno 2009: a SWOT analysis. *Tech. Digest IEEE Electron Devices Meet.* 461–464 (IEEE, 2009).
37. Kamata, Y. High-*k*/Ge MOSFETs for future nanoelectronics. *Mater. Today* **11,** 30–38 (2008).
38. Xie, Y. H. *et al.* Very high mobility two-dimensional hole gas in Si/Ge_xSi_{1-x}/Ge structures grown by molecular beam epitaxy. *Appl. Phys. Lett.* **63,** 2263–2264 (1993).
39. Engelhardt, C. M. *et al.* High mobility 2-D hole gases in strained Ge channels on Si substrates studied by magnetotransport and cyclotron resonance. *Solid State Electron.* **37,** 949–952 (1994).
40. Madhavi, S., Venkataraman, V. & Xie, Y. H. High room temperature hole mobility in $Ge_{0.7}Si_{0.3}$/Ge/$Ge_{0.7}Si_{0.3}$ modulation doped heterostructures in the absence of parallel conduction. *J. Appl. Phys.* **89,** 2497–2499 (2001).
41. Irisawa, T., Miura, H., Ueno, T. & Shiraki, Y. Channel width dependence of mobility in Ge channel modulation doped structures. *Jpn. J. Appl. Phys.* **40,** 2694–2696 (2001).
42. Koester, S. J., Hammond, R. & Chu, J. O. Extremely high transconductance Ge/$Si_{0.4}Ge_{0.6}$ p-MODFET's grown by UHV-CVD. *IEEE Electron Device Lett.* **21,** 110–112 (2000).
43. Houssa, M. *et al.* Ge dangling bonds at the (100)Ge/GeO_2 interface and the viscoelastic properties of GeO_2. *Appl. Phys. Lett.* **93,** 161909 (2008).
44. Atalla, M. M., Tannenbaum, E. & Scheibner, E. J. Stabilization of silicon surfaces by thermally grown oxides. *Bell Syst. Tech. J.* **38,** 749–783 (1959).
45. Kahng, D. Silicon–silicon dioxide surface devices. Technical memorandum (Bell Laboratories, 16 January 1961); reprinted in Sze, S. M. (ed.) *Semiconductor Devices: Pioneering Papers* 583–596 (World Scientific Publishing, 1991).
46. Currie, M. T., Samavedam, S. B., Langdo, T. A., Leitz, C. W. & Fitzgerald, E. A. Controlling threading dislocation densities in Ge on Si using graded SiGe layers and chemical–mechanical polishing. *Appl. Phys. Lett.* **72,** 1718–1720 (1998).
47. Park, J. S. *et al.* Defect reduction of selective Ge epitaxy in trenches on Si (001) substrates using aspect ratio trapping. *Appl. Phys. Lett.* **90,** 052113 (2007).
48. Wang, G. *et al.* Fabrication of high quality Ge virtual substrates by selective epitaxial growth in shallow trench isolated Si (001) trenches. *Thin Solid Films* **518,** 2538–2541 (2010).
49. Wang, G. *et al.* High quality Ge epitaxial layers in narrow channels on Si (001) substrates. *Appl. Phys. Lett.* **96,** 111903 (2010).
50. Taraschi, G., Pitera, A. J. & Fitzgerald, E. A. Strained Si, SiGe, and Ge on-insulator: review of wafer bonding fabrication techniques. *Solid State Electron.* **47,** 1297–1305 (2004).
51. Nakaharai, S., Tezuka, T., Sugiyama, N., Moriyama, Y. & and Takagi, S. Characterization of 7-nm-thick strained Ge-on-insulator layer fabricated by Ge-condensation technique. *Appl. Phys. Lett.* **83,** 3516–3518 (2003).
52. Hashimoto, T., Yoshimoto, C., Hosoi, T., Shimura, T. & Watanabe, H. Fabrication of local Ge-on-insulator structures by lateral liquid-phase epitaxy: effect of controlling interface energy between Ge and insulators on lateral epitaxial growth. *Appl. Phys. Express* **2,** 066502 (2009).

Author Information
Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature.
Correspondence should be addressed to the author (ravi.pillarisetty@intel.com).

doi:10.1038/nature10679

Tunnel field-effect transistors as energy-efficient electronic switches

Adrian M. Ionescu[1] & Heike Riel[2]

Power dissipation is a fundamental problem for nanoelectronic circuits. Scaling the supply voltage reduces the energy needed for switching, but the field-effect transistors (FETs) in today's integrated circuits require at least 60 mV of gate voltage to increase the current by one order of magnitude at room temperature. Tunnel FETs avoid this limit by using quantum-mechanical band-to-band tunnelling, rather than thermal injection, to inject charge carriers into the device channel. Tunnel FETs based on ultrathin semiconducting films or nanowires could achieve a 100-fold power reduction over complementary metal–oxide–semiconductor (CMOS) transistors, so integrating tunnel FETs with CMOS technology could improve low-power integrated circuits.

Reducing the size of complementary metal–oxide–semiconductor (CMOS) field-effect transistors (FETs) has enabled extraordinary improvements in the switching speed, density, functionality and cost of microprocessors. But advanced CMOS technology now faces two problems[1] that together result in high power consumption: the increasing difficulty in further reducing the supply voltage, and stopping the rising leakage currents that degrade the switching ratio of 'on' and 'off' currents (I_{ON}/I_{OFF}). Recent reviews[2,3] have highlighted the need for new devices that can compete with or complement CMOS transistors.

Here we review the physics, design and optimization of one such device, the tunnel FET (TFET), and consider the potential and drawbacks of this energy-efficient device that could bring the voltage supply of integrated circuits below 0.5 V. We discuss the technology boosters needed to increase its performance, the reduced sensitivity of its direct current characteristics to gate length, and the variability of its electrical characteristics to changes in conditions. Finally, we compare its performance in various material systems — silicon, group III–V compounds and carbon — and discuss the related problems.

The quest for an energy-efficient switch

In a metal–oxide–semiconductor FET (MOSFET), the current-switching process involves the thermionic (temperature-dependent) injection of electrons[4,5] over an energy barrier. This sets a fundamental limit to the steepness of the transition slope from the off to the on state. The gate voltage required to change the drain current by one order of magnitude when the transistor is operated in the subthreshold region is reflected in the expression of the subthreshold swing, S:

$$S = \underbrace{\frac{dV_G}{d\Psi_S}}_{m}\underbrace{\frac{d\Psi_S}{d(\log_{10} I_D)}}_{n} \cong \left(1 + \frac{C_d}{C_{ox}}\right)\ln 10\,\frac{kT}{q} \rightarrow \frac{kT}{q}\ln 10 \cong 60\ \text{mV decade}^{-1}\,|\,T = 300\ \text{K} \qquad (1)$$

where V_G is the gate voltage, I_D is the drain current, kT/q is the thermal voltage, and C_d and C_{ox} are the depletion and the oxide capacitances, respectively. The term m is the transistor body factor, and n is a factor that characterizes the change of the drain current with the surface potential, Ψ_S, reflecting the conduction mechanism in the channel. A subthermal S would be less than kT/q ln10 and could be obtained by using new physical principles rather than thermionic injection.

As the transistor gate length is reduced, improved performance requires the supply voltage, V_{DD}, and simultaneously the threshold voltage, V_T, to be lowered to keep the overdrive factor ($V_{DD} - V_T$) high. As a consequence, the leakage current, I_{OFF}, increases exponentially (see the vertical intercept of I–V plots in Fig. 1a) because the S of a MOSFET is not scalable but has a minimum value of 60 mV per decade (that is, it takes 60 mV to increase the current by one order of magnitude) at room temperature. Typical values of S in advanced CMOS technology are close to 100 mV per decade; by lowering V_{DD} from 500 mV to 250 mV while preserving the overdrive, the leakage power has been shown to increase unacceptably by a factor of 275 in a 45-nm bulk CMOS technology[6].

Another way of reducing the voltage supply without performance loss is to increase the turn-on steepness, which means decreasing the average subthreshold swing, S_{avg}[7,8], defined as:

$$S_{avg} = \frac{V_T - V_{GOFF}}{\log \dfrac{I_T}{I_{OFF}}} \approx \frac{V_{DD}}{\log \dfrac{I_{ON}}{I_{OFF}}} \qquad (2)$$

Therefore, devices with a steep S, called steep-slope switches, are expected to enable V_{DD} scaling.

Figure 1b shows a qualitative comparison of some major candidates to improve the characteristics of bulk silicon MOSFET switches: multigate devices for improved electrostatics; high-mobility channels exploiting group III–V and SiGe materials; and TFETs that use quantum-mechanical tunnelling. At moderate performance requirements, such as operation point A, TFETs offer not only improved I_{ON}/I_{OFF}, but also superior performance (higher I_{ON} at the same voltage) or power savings at the same performance (lower voltage for the same I_{ON}) over MOSFETs. However, when a much higher performance is required, such as at operation point B, a MOSFET is the better solution.

The energy efficiency of a logic operation can be evaluated by analysing its switching energy diagram[9,10] (Fig. 1c), showing the balance of the dynamic, $E_{dynamic}$, and the leakage, $E_{leakage}$, components of the total switching energy, E, versus the V_{DD}:

$$\begin{aligned} E_{total} &= E_{dynamic} + E_{leakage} = \alpha L_d C V_{DD}^2 + L_d I_{OFF} V_{DD} \tau_{delay} \\ &\approx \alpha L_d C V_{DD}^2 = L_d C V_{DD}^2 \frac{I_{OFF}}{I_{ON}} = L_d C V_{DD}^2 \left(\alpha + \frac{I_{OFF}}{I_{ON}}\right) \\ &\approx L_d C V_{DD}^2 \left(\alpha + 10^{\frac{-V_{DD}}{S}}\right) \end{aligned} \qquad (3)$$

[1]Ecole Polytechnique Fédérale Lausanne, 1015 Lausanne, Switzerland. [2]IBM Research – Zurich, 8803 Rüschlikon, Switzerland.

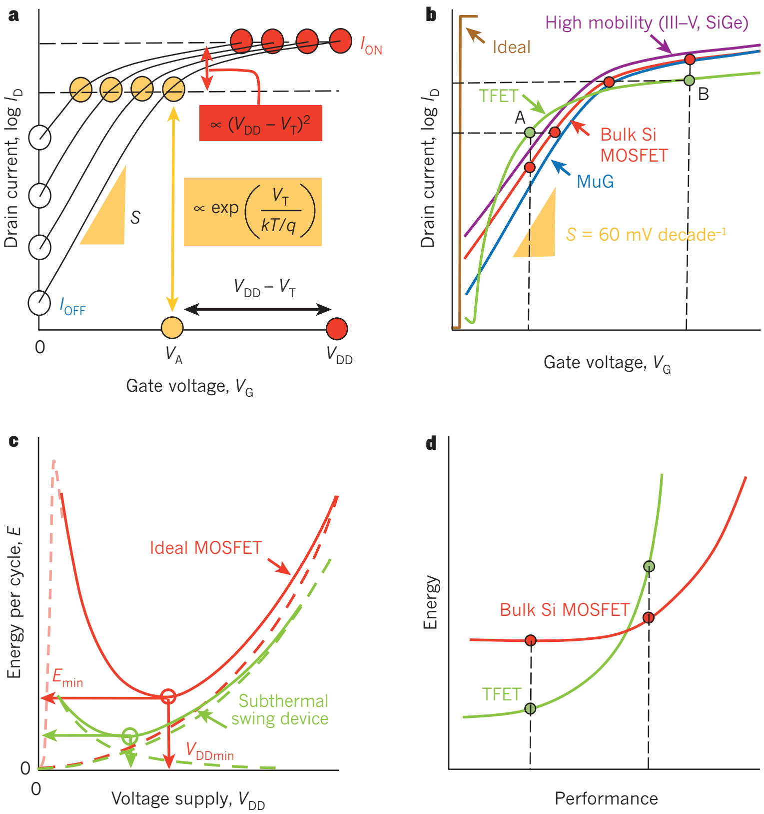


Figure 1 | Power challenge and main characteristics of an energy-efficient tunnel FET. **a**, Transfer characteristics (drain current, I_D, and gate voltage, V_G) of a MOSFET switch showing an exponential increase in I_{OFF} (more than tenfold increase for every 60 mV at room temperature) because of an incompressible subthreshold swing, *S*. Here the simultaneous scaling down of both the supply voltage, V_{DD}, and the threshold voltage, V_T, maintains the same performance (I_{ON}) by keeping the overdrive ($V_{DD} - V_T$) constant. **b**, Qualitative comparison of three engineering solutions to improve the characteristics of the bulk silicon MOSFET switch (red): a multigate device (MuG, blue) for improved electrostatics; a high-mobility channel (purple) using group III–V and SiGe materials; and a TFET (green), which has a steep off–on transition and the lowest I_{OFF}. At operation point A, because of its subthermal subthreshold swing, the TFET offers not only an improved I_{ON}/I_{OFF} but also a superior performance and a power saving at the same performance as a MOSFET. At operation point B, corresponding to higher performance, the MOSFET switch becomes the better solution. **c**, Comparison of the minimum switching energy, E_{min}, and the corresponding voltage supply, V_{DDmin}, for a subthermal swing device ($S < 60$ mV decade^{-1}, green curve) and the ideal MOSFET ($S = 60$ mV decade^{-1}, red) at the same I_{ON}/I_{OFF}. **d**, Comparison between switching energy and performance for a MOSFET and a TFET. The steep-swing TFET offers better energy efficiency at lower or moderate performance level.

where L_d is the logic depth, *C* is the switched capacitance, τ_{delay} is the delay time and α is the logic activity factor (typically ~0.01). The operation frequency, *f*, can be expressed as

$$f = \frac{1}{L_d \tau_{delay}} \qquad (4)$$

and in modern technologies can be considered empirically as being proportional to V_{DD} (ref. 11). Therefore the power dissipation, *P*, is

$$P = \alpha L_d C V_{DD}^2 f = I_{OFF} V_{DD} \approx KCV_{DD} = I_{OFF} V_{DD}^3 \qquad (5)$$

Consequently, a technology that would enable a fivefold voltage scaling (from 1.0 V to 0.2 V) with a negligible leakage power could offer a 125-fold power dissipation reduction. From equation (3), it seems that CMOS logic has a lower limit in energy per operation, E_{min}, owing to the exponential increase of the subthreshold leakage, I_{OFF}, with V_{DD} scaling (see Fig. 1c). Hanson *et al.*[10] showed that E_{min} is proportional to the switched capacitance multiplied by the square of the *S*, $C \times S^2$, whereas V_{DDmin} is proportional to *S*. Nose and Sakurai[12] demonstrated that for an optimized CMOS circuit design, the ratio of leakage to dynamic energy is approximately 0.3–0.5 across a wide range of parameters.

Equation (3) shows that at the same performance (I_{ON} and *f*), any device that can offer the required I_{ON}/I_{OFF} at a lower V_{DD} (based on a smaller *S*) will always be more energy efficient (lower E_{min} and V_{DDmin}). This is depicted in Fig. 1d, which compares the switching energy for a TFET and a MOSFET as a function of the performance required.

Many device innovations to lower *S* below the MOSFET thermal limit, by decreasing the factors *m* and *n* in equation (1), have been proposed. Reducing *n* to achieve a subthermal *S* involves a modification of the carrier-injection mechanism. For this, impact ionization[13] and quantum-mechanical band-to-band tunnelling (BTBT)[14] in TFETs have been proposed. Another alternative to decrease *S* is to reduce the body factor, *m*, to a value smaller than 1. This can be achieved by using the recently proposed negative-capacitance FET (NC-FET)[15–17] or micro- or nano-electromechanical (M/NEM) movable electrodes in M/NEM-FET or NEM relay devices[18–20], in which the instability points between the electrical and the mechanical force are used to define super-abrupt transitions between the off and on states. Experimentally, an *S* of less than 2 mV per decade has been demonstrated[18], but electromechanical devices have their own limitations, such as voltage-scaling limitations, reliability issues and a stringent need for a controlled environment for robust operation.

In this review, we concentrate on the TFET. The gated p-i-n structure, comprising a p- and an n-doped region on either side of a gated intrinsic region, was proposed in 1978 by Quinn *et al.*[21]. Banerjee *et al.*[22] studied the behaviour of a three-terminal silicon TFET, and Takeda *et al.*[23] explored various aspects specifically related to the scaling down. Baba[24] fabricated TFETs called surface tunnel transistors in group III–V materials. In 1995, Reddick and Amaratunga[25] reported experiments on silicon surface tunnel transistors. In 1996, Koga and Toriumi[26] proposed a three-terminal 'forward-biased' silicon tunnelling device as a post-CMOS switch candidate. In 2000, Hansch *et al.*[27] published experimental results on a reverse-biased vertical silicon TFET that had a highly doped boron delta-layer fabricated by molecular beam epitaxy. Aydin *et al.*[28] processed lateral TFETs on silicon-on-insulator (SOI) in 2004, which in principle were similar to TFETs without an intrinsic region. The gate over a p–n junction was intended to reduce the gate capacitance to increase the speed. Recently, TFETs fabricated in various material systems (carbon, silicon, SiGe and group III–V materials)[29–33] have emerged experimentally as the most promising candidates for switches with ultralow standby power and sub-0.5 V logic operation.

The physics of TFETs

In contrast to MOSFETs, in which charge carriers are thermally injected over a barrier, the primary injection mechanism in a TFET is interband tunnelling, whereby charge carriers transfer from one energy band into another at a heavily doped p^+–n^+ junction. This tunnelling mechanism was first identified by Zener[14] in 1934. In a TFET, interband tunnelling can be switched on and off abruptly by controlling the band bending in the channel region by means of the gate bias. This function can be realized in a reverse-biased p-i-n structure (Fig. 2a). In principle, the TFET

is an ambipolar device, showing p-type behaviour with dominant hole conduction and n-type behaviour with dominant electron conduction.

However, by designing an asymmetry in the doping level or profile, or by restricting the movement of one type of charge carrier using heterostructures, the tunnelling barrier at the drain can be widened to suppress the ambipolarity[8,34]. The asymmetry also achieves a low off-state current[3]. In the TFET off state (dashed blue line in Fig. 2b), the valence band edge of the channel is located below the conduction band edge of the source, so BTBT is suppressed, leading to very small TFET off-state currents that are dictated by the reverse-biased p-i-n diode. Applying a negative gate voltage (solid red curve in Fig. 2b) pulls the energy bands up. A conductive channel opens as soon as the channel valence band has been lifted above the source conduction band because carriers can now tunnel into empty states of the channel. Because only carriers in the energy window $\Delta\Phi$ can tunnel into the channel, the energy distribution of carriers from the source is limited; the high-energy part of the source Fermi distribution is effectively cut off[35], as shown in Fig. 2b. Thus the electronic system is effectively 'cooled down', acting as a conventional MOSFET at a lower temperature. This filtering function makes it possible to achieve an S of below 60 mV per decade (Fig. 2c). However, the channel valence band can be lifted by a small change in gate voltage, and the tunnelling width can effectively be reduced by the gate voltage[36,37]. As a consequence of the BTBT mechanism, S in a TFET is not constant, but depends on the applied gate–source bias, as indicated in Fig. 2c, increasing with the gate-to-source bias. The key to the better voltage scaling of a TFET than a MOSFET is that S remains below 60 mV per decade over several orders of magnitude of drain current.

One challenge in TFETs is to realize high on currents because I_{ON} critically depends on the transmission probability, T_{WKB}, of the interband tunnelling barrier. This barrier can be approximated by a triangular potential, as indicated by the grey shading in Fig. 2b, so T can be calculated using the Wentzel–Kramer–Brillouin (WKB) approximation[4,36]:

$$T_{\mathrm{WKB}} \approx \exp\left(-\frac{4\lambda\sqrt{2m^*}\sqrt{E_g^{\,3}}}{3q\hbar\,(E_g+\Delta\Phi)}\right) \qquad (6)$$

where m^* is the effective mass and E_g is the bandgap. Here, λ is the screening tunnelling length and describes the spatial extent of the transition region at the source–channel interface (Fig. 2b); it depends on the specific device geometry. In a TFET, at constant drain voltage, V_D, the V_G increase reduces λ and increases the energetic difference between the conduction band in the source and the valence band in the channel ($\Delta\Phi$), so that in a first approximation the drain current is a super-exponential function[37] of V_G. As a result, in contrast to the MOSFET, the point subthreshold swing of the TFET is no longer a constant but strongly depends on V_G. The smallest subthermal values occur at the lowest gate voltages. A high on current requires a high transparency of the tunnelling barrier, thus maximizing T_{WKB}, which in the best case should be unity. Equation (6) suggests optimized design approaches to boost the on current. Luisier and Klimeck[38] found that the WKB approximation works properly in direct bandgap semiconductors, such as InAs (if one single imaginary path connecting the valence band and the conduction band dominates the tunnelling process), but has limited accuracy for Si and Ge structures or when quantum effects and phonon-assisted tunnelling become dominant.

Fundamental performance boosters

The goals for TFET optimization are to simultaneously achieve the highest possible I_{ON}, the lowest S_{avg} over many orders of magnitude of drain current, and the lowest possible I_{OFF}. To outperform CMOS transistors, the target parameters for TFETs are: I_{ON} in the range of hundreds of milliamperes; S_{avg} far below 60 mV per decade for five decades of current; $I_{ON}/I_{OFF} > 10^5$; and $V_{DD} < 0.5$ V. Because S decreases with the V_G (Fig. 2c), TFETs are naturally optimized for low-voltage operation.

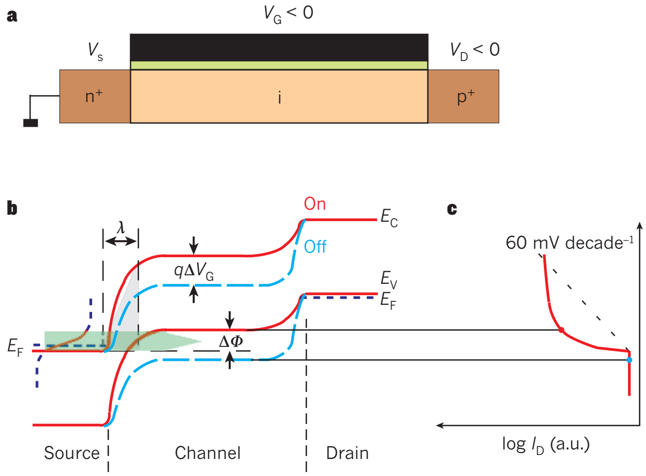


Figure 2 | Principle of operation of a TFET. **a**, Schematic cross-section of p-type TFET with applied source (V_S), gate (V_G) and drain (V_D) voltages. **b**, Schematic energy band profile for the off state (dashed blue lines) and the on state (red lines) in a p-type TFET. In the off state, no empty states are available in the channel for tunnelling from the source, so the off current is very low. Decreasing V_G moves the valence band energy (E_V) of the channel above the conduction band energy (E_C) of the source so that interband tunnelling can occur. This switches the device to the on state, in which electrons in the energy window, $\Delta\Phi$ (green shading), can tunnel from the source conduction band into the channel valence band. Electrons in the tail of the Fermi distribution cannot tunnel because no empty states are available in the channel at their energy (dotted black line), so a slope of less than 60 mV decade^{-1} can be achieved. This is indicated in the schematic transfer characteristics shown in **c**. In contrast to a conventional MOSFET, a TFET has a slope that is not linear on a logarithmic scale, which can be explained by the complex dependency of the tunnel current on the transmission probability through the barrier, as well as on the number of available states determined by the source and channel Fermi functions. The BTBT can be approximated by the triangular potential barrier indicated in grey. Because the tunnel current depends on the transmission probability through the barrier, as well as on the number of available states determined by the source and channel Fermi functions, the resultant slope is not linear on a logarithmic scale, which it is for a conventional MOSFET. λ, screening tunnelling length. a.u., arbitrary units; E_F, Fermi energy.

To realize a high tunnelling current and a steep slope, the transmission probability of the source tunnelling barrier should become close to unity for a small change in V_G. The WKB approximation, shown in equation (6), suggests that the bandgap (E_g), the effective carrier mass (m^*) and the screening tunnelling length (λ) should be minimized for high barrier transparency. Whereas E_g and m^* depend solely on the material system, λ is strongly influenced by several parameters, such as the device geometry, dimensions, doping profiles and gate capacitance[3,39,40]. A small λ results in a strong modulation of the channel bands by the gate. This requires a high-permittivity (high-κ) gate dielectric[8] with as low an equivalent oxide thickness as possible. Furthermore, the body thickness of the channel should be minimized, showing in the best case one-dimensional electronic transport behaviour[36,39]. The abruptness of the doping profile at the tunnel junction is also important. To minimize the tunnelling barrier, the high source doping level must fall off to the intrinsic channel in as short a width as possible. This requires a change in the doping concentration of about 4–5 orders of magnitude within a distance of only a few nanometres[40,41]. Increasing the source doping reduces λ and may lead to a slightly smaller energy barrier at the tunnel junction because of bandgap narrowing[40]. However, the energy filtering effect described above becomes effective only if the Fermi energy in the source is not too large[36].

TFETs do not follow the same scaling rules as MOSFETs, in which many parameters must be scaled simultaneously to keep the same electric field throughout the device[42]. In a TFET, the high electric fields exist only at the junctions. The current is determined by the screening

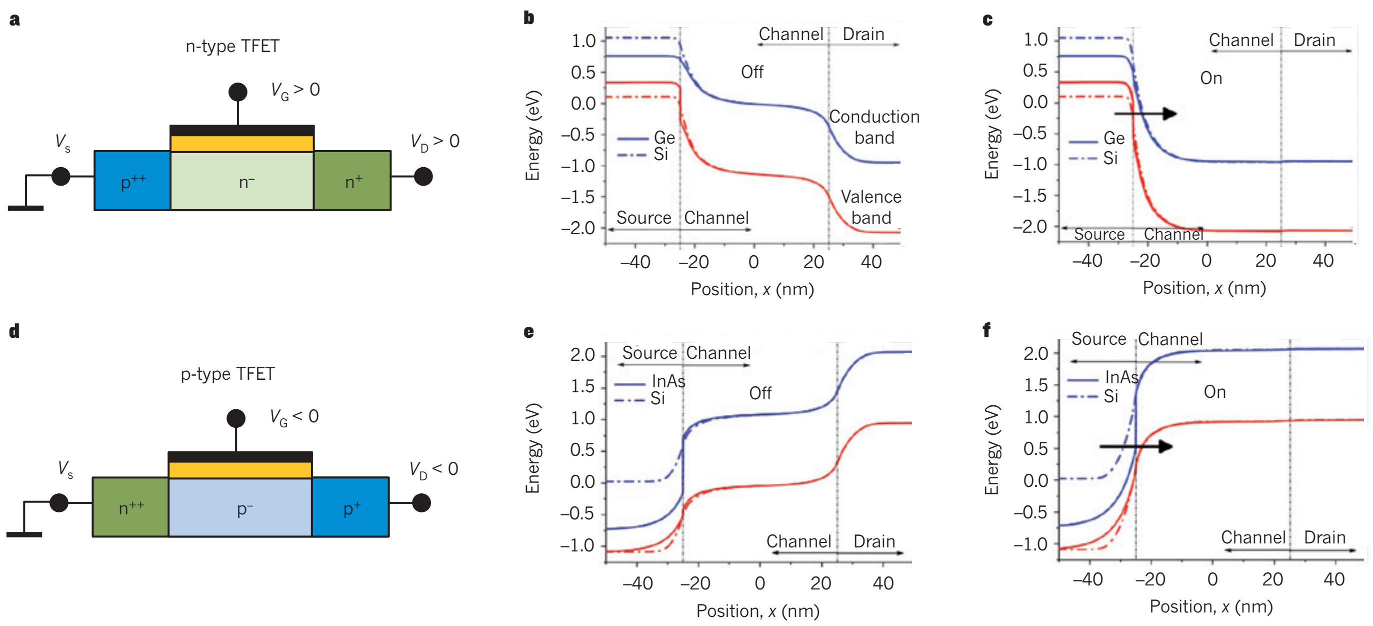


Figure 3 | Band diagrams of heterostructure C-TFETs. C- TFET device n-type (**a**–**c**) and p-type (**d**–**f**) architectures (**a**, **b**) and related band diagrams in the off (**b**, **e**) and on (**c**, **f**) state for two major implementations: all-silicon n- and p-type devices (dashed lines) and heterostructures with a Ge source and Si channel for the n-type switch and an InAs source and silicon channel for the p-type switch (solid lines). Band diagrams correspond to device architectures with a channel length of 50 nm and a high-κ dielectric thickness of 3 nm. The graphs show the conduction band (blue) and the valence band (red).

tunnelling length, so that the length of the intrinsic region has little effect on the device characteristics, as long as the length is above some critical length, L_{crit} (~20 nm for silicon TFETs[42]), at which p-i-n leakage becomes predominant. Experimental results[43] confirmed the lack of dependence of I_{ON} on TFET length.

Device optimization should apply to both n- and p-type TFETs simultaneously, to offer a complementary TFET (C-TFET) technology for logic circuits. In a heterostructure TFET, the materials are chosen such that the source material has a small bandgap, so that the width of the energy barrier at the source junction is reduced in the on state, whereas the drain material has a large bandgap, which creates the largest possible energy barrier width at the drain side in the off state to keep the off current low. The way in which the bands line up with each other at the heterojunction is also crucial[44,45]. Knoch[46] suggested that although a broken line-up yields the best I_{ON}, only $S \geq 60$ mV per decade is obtained; therefore, a combination of steep S and high I_{ON} can been achieved with moderate doping and a staggered band lineup ($S = 33$ mV per decade and an intrinsic cutoff frequency of 4.74 THz)[47]. Koswatta *et al.*[48] included electron–phonon scattering in the transport model and found the best TFET performance for a broken-gap heterojunction.

Another optimization criterion is maximizing the gate modulation of the tunnelling barrier width by an appropriate alignment of the tunnelling path with the direction of the electric field modulated by the gate. By overlapping the gate with the tunnelling region, or designing a source region covered with an epitaxial intrinsic channel layer under the top gate, I_{ON} can be improved by a factor of more than 10 and a low S_{avg} can be obtained[49–52].

The effect of a staggered bandgap at the TFET source-to-channel junction as a technology booster is shown in Fig. 3. The aim is to find a solution for an optimized C-TFET. The simulated devices have a silicon channel length of 50 nm, a 20-nm-thick film and asymmetric

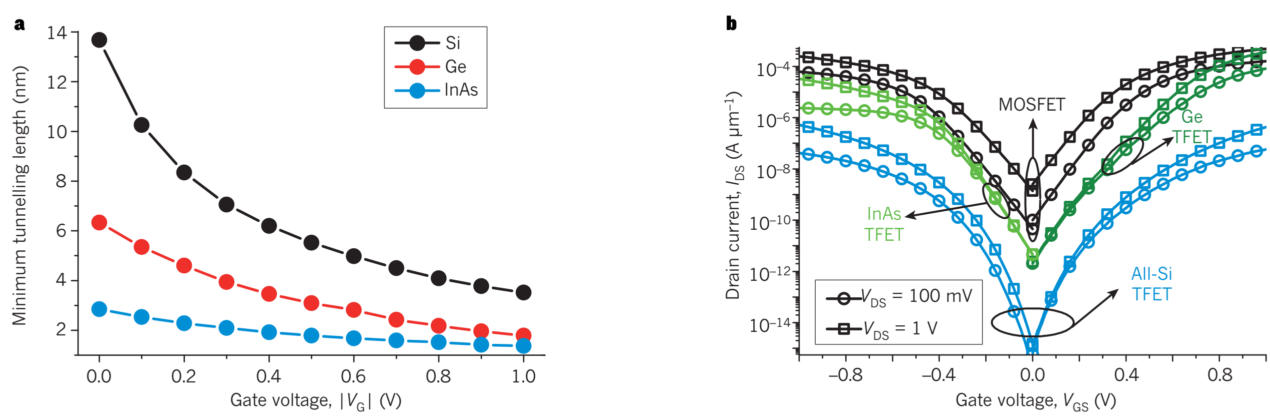


Figure 4 | Importance of the material system on TFET performance. **a**, Modulation of the minimum screening tunnelling length with the applied gate voltage in all-silicon (black), Ge-source (red) and InAs-source (blue) TFETs, showing the beneficial effect of a higher tunnelling rate due to the shorter tunnelling length in a heterostructure TFET with a low bandgap source material compared with silicon. By contrast, a higher ratio between the tunnelling length in the off and the on state reflects an improved I_{ON}/I_{OFF}. **b**, Corresponding transfer characteristics of a state-of-the-art 65-nm CMOS transistor (black), complementary Ge/InAs TFET (green) and complementary all-Si TFET (blue). The complementary Ge/InAs TFET achieves the best trade-off between a low I_{OFF}, a steep subthreshold swing and performance. I_{DS}, drain-to-source current; V_{DS}, drain-to-source voltage; V_{GS}, gate voltage at source.

doping to avoid ambipolarity. The cross-section and bands in the off and the on state of the n-type TFET are shown in Fig. 3a–c. Changing the source material from Si to Ge considerably improves λ and $\Delta\Phi$ at the same applied V_G in the on state. A similar band-engineering optimization performed for the p-type device with an InAs source is shown in Fig. 3d–f. The corresponding reduction of the screening tunnelling length and its dependence on the V_G when the source bandgap is changed is depicted in Fig. 4a. The simulated I_D versus V_G characteristics of a 65-nm CMOS technology node with 50-nm all-Si C-TFETs and 50-nm Ge/InAs C-TFETs are compared in Fig. 4b. At $V_D = V_G = 1$ V, the Ge and InAs TFETs have on currents of 244 μA μm^{-1} and 83 mA mm^{-1}, respectively, which means improvements by factors of 480 and 162, respectively, over their all-Si TFET counterparts, and much lower I_{OFF} and I_{ON}/I_{OFF} than the 65-nm CMOS device. Their average swing over three decades of drain current is close to 60 mV per decade, showing that further optimization is needed. Heterostructure TFETs[33,45,46,53] similar to those reported here can offer viable solutions for on currents higher than 100 μA μm^{-1}, $I_{ON}/I_{OFF} > 10^7$ and V_{DD} smaller than 0.5 V. The all-Si C-TFETs have the lowest I_{OFF} and average subthreshold swings of less than 40 mV per decade, but their I_{ON} is not a good trade-off for performance compared with CMOS transistors.

Band-to-band tunnelling

The fundamental performance boosters of TFETs described before require engineering solutions concerning their design, choice of materials and integration with advanced silicon platforms. In this section we discuss existing and state-of-the-art research efforts aimed at TFET optimization, design and implementation, together with their experimental or predicted electrical performance.

All-silicon TFETs

TFETs offer the potential for a low off current and a small S, but they generally have a lower on current than conventional MOSFETs, so a smart design strategy could achieve a small S_{avg} and a high I_{ON} without degrading I_{OFF}. As with CMOS technology boosters, performance boosters for TFETs should not be suggested independently. Instead, an additive strategy of boosters for the same device should be applied, so that improvements in device performance are cumulative[54]. The major technology boosters for all-silicon TFETs include[54,55] the use of a high-κ gate dielectric, a more abrupt doping profile at the tunnel junction, a thinner body, higher source doping, a double gate, a gate oxide aligned with the intrinsic region, and a shorter intrinsic region (and gate) length.

The physics of TFETs are governed by the BTBT rate, so they differ from those of conventional MOSFETs. It is therefore likely that the sensitivity of the device's characteristics to variations in the technology parameters will differ too, which can imply new technology challenges. Boucart[56] predicted that TFET performance will be much less sensitive to doping fluctuations and gate length scaling than in conventional CMOS transistors. By contrast, control of the high-κ gate process, the abruptness of doping at the tunnel junction, and the film thickness in ultrathin-body SOI devices, with significantly less parameter variation than that required by CMOS devices, is crucial for building future TFETs with reproducible characteristics.

Figure 5a shows a recent all-silicon lateral TFET fabricated on fully depleted silicon with a silicon film thickness of 20 nm and a gate length of 100 nm[31,57] that benefits from some major technology boosters and from advanced engineering of the contacts. The device includes a high-κ gate stack (3 nm HfO_2, and 3 nm and 10 nm TiN), a double epitaxially raised Si source–drain, and an asymmetric n^+ source and p^+ drain achieved in two successive lithography and implantation steps. This all-silicon TFET shows extremely low I_{OFF} values (~10–100 fA μm^{-1}; Fig. 5b), but I_{ON} is less than 0.1 μA mm^{-1} at a V_{DD} of 1 V. By applying the same process and design to $Si_{1-x}Ge_xOI$ substrates ($x = 0$, 15% or 30%), extraordinary improvements in I_{ON} can be obtained[57]: a 335-fold improvement for the n-TFET and a 2,700-fold improvement for the p-TFET, compared with the SOI TFET counterparts (Fig. 5c). In recent

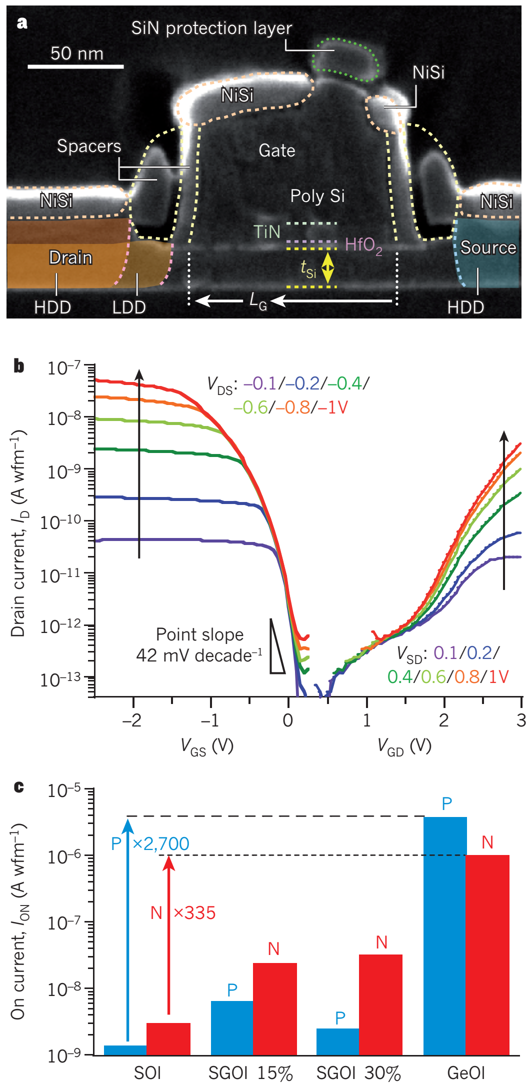


Figure 5 | Implementation of all-Si technology boosters. **a**, Scanning electron micrograph showing the cross-section of an SOI TFET from CEA-Leti that implemented some major technology boosters. Image reprinted, with permission, from ref. 31. **b**, I_D–V_G characteristics of the all-silicon TFET, showing a subthreshold swing at room temperature of 42 mV decade^{-1} and an I_{OFF} smaller than 100 fA μm^{-1} at $V_{DS} = 1$ V (different colours correspond to different values of V_D). **c**, A major improvement in I_{ON} obtained in an all-silicon C-TFET by applying the same design and fabrication of TFET on $Si_{1-x}Ge_xOI$ (molar fraction $x = 0$, 15% and 30%). Note the 335-fold and 2,700-fold improvements for the n-TFET (N) and p-TFET (P), respectively, for the SiGeOI over the SOI TFET. V_{DS}, drain-to-source voltage; V_{GD}, gate voltage at drain; V_{GS}, gate voltage at source; V_{SD}, source-to-drain voltage.

optimizations of an all-silicon TFET[58], values of I_{ON} close to 100 μA μm^{-1} in sub-60-nm devices were obtained.

Group III–V-semiconductor-based TFETs

A further strategy to improve I_{ON} and S is to use low-bandgap and low-effective-mass materials and band engineering to increase BTBT. For this, group III–V materials are very attractive as they can provide small tunnelling mass and allow different band-edge alignments.

Simulation showed that by reducing only the bandgap of the TFET material from Si to InAs or InSb, the I_{ON} increases by several orders of magnitude and can be reached at lower electric fields[3]. Recent experimental results for InGaAs TFETs indicate that a higher I_{ON} at a lower V_G than with Si TFETs seems possible[59,60]. The first InGaAs TFET by Mookerjea *et al.*[59] achieved an on current of 20 μA μm^{-1} with an S of

250 mV per decade, whereas Zhao *et al.*[60] improved I_{ON} to 50 μA μm^{-1} with an S of around 90 mV per decade, which is the best local swing achieved so far for III–V-based TFETs but is still above the thermal limit of MOSFETs. The degraded S is attributed to parasitic tunnelling mechanisms involving traps in the source tunnel junction[61].

The effective bandgap for tunnelling can be decreased even further by using heterostructures. Although there is not yet full agreement on whether a staggered or a broken gap alignment works best, all theoretical studies predict that the TFET performance can be significantly enhanced compared with homojunctions[45–47,62]. The first experimental implementations of III–V heterojunction TFETs have only recently been demonstrated[63,64]. To reduce the tunnelling barrier, InAs and GaAsSb were chosen for the source, with AlGaSb and InGaAs for the channel. Another reason for selecting these materials is that they allow lattice-matched growth, and thus the use of conventional III–V growth and processing technologies.

For the C-TFET technology, the combination of an InAs source with a Si drain and channel yields the best on-state performance for the p-TFET[45]. However, InAs and Si possess a lattice mismatch of about 11%, which results in highly defective material growth and prevents integration onto Si in a conventional approach. This challenge can be met by using grown nanowires. They are attractive materials as they can be grown epitaxially via metal–organic chemical vapour deposition directly on Si(111) (refs 65, 66).

A comprehensive study has experimentally investigated[67] the quality and suitability of the InAs–Si heterojunction, which is a key element for high-performance TFETs. In particular, studying highly doped p–n junctions (so-called Esaki tunnel diodes) provides insight into the tunnel process, and thus the limits of the TFET drive can be explored. Figure 6a shows a schematic cross-section of such an InAs–Si heterostructure nanowire device in which the tunnel junction is located between the n-type InAs nanowire and the p-type Si substrate. The highest Si doping of 10^{20} atoms cm^{-3} produced tunnel diodes with current densities as high as 250 kA cm^{-2} at 0.5 V reverse bias (see Fig. 6c). The high tunnel currents and the observed negative differential resistance are indications of a well-defined and abrupt Si–InAs heterojunction[66].

Moreover, the vertical nanowire provides an optimal geometry for minimizing the screening tunnelling length, λ, by using a 'gate-all-around' (GAA) architecture, in which the gate is wrapped around the cylindrical nanowire channel to provide the best electrostatic control. Simulations have shown that λ is reduced three- to fourfold when using the GAA nanowire architecture compared with using a single-gate device with a channel thickness or diameter of 10 nm[68]. The cross-section of a recently fabricated vertical n-i-p InAs–Si–Si nanowire heterojunction TFET with InAs as a low-bandgap source[69] is shown in Fig. 6b. The single-nanowire TFET exhibits switching operation under reverse-bias conditions with an S of ~220 mV per decade and a drive current of ~0.4 μA μm^{-1} (see Fig. 6d). This first InAs-source, Si-channel TFET implementation needs further optimization. It is anticipated that the drive current can be improved by introducing n-type doping in the InAs wire, by scaling the equivalent oxide thickness, and by reducing the contact resistance.

Carbon-based TFETs

Carbon nanotubes and graphene nanoribbons are appealing materials for use in TFETs. The light effective mass of their charge carriers, their small and direct bandgap, and their excellent electrostatic control of the gate over the channel owing to the ultrathin body make them among the best choices in terms of both materials and device geometry. The first ever TFET demonstrating an S of less than 60 mV per decade was achieved with a carbon nanotube structure by Appenzeller *et al.*[29] in 2004. In their device, the electrostatics in the carbon nanotube were controlled by two independent gates. A common back gate electrostatically doped both the source and drain, and another gate allowed control of the channel bands, creating a p-i-p FET device. Although the on current was low because the carriers had to tunnel through two barriers, an S of 40 mV per decade was achieved for the first time. Furthermore, simulation and temperature-dependent measurements provided clear evidence that a carbon nanotube TFET had indeed been realized[29,35]. More advantageous would be the implementation of a p-i-n-structured carbon nanotube TFET. Progress has been hampered by the difficulty of establishing appropriately doped carbon nanotube regions and abrupt junctions. So far there has been little experimental verification of carbon nanotube TFETs, but several theoretical studies have been conducted[70]. The influence of electron–phonon scattering on the performance of carbon nanotube TFETs has also been investigated[71], and there is strong

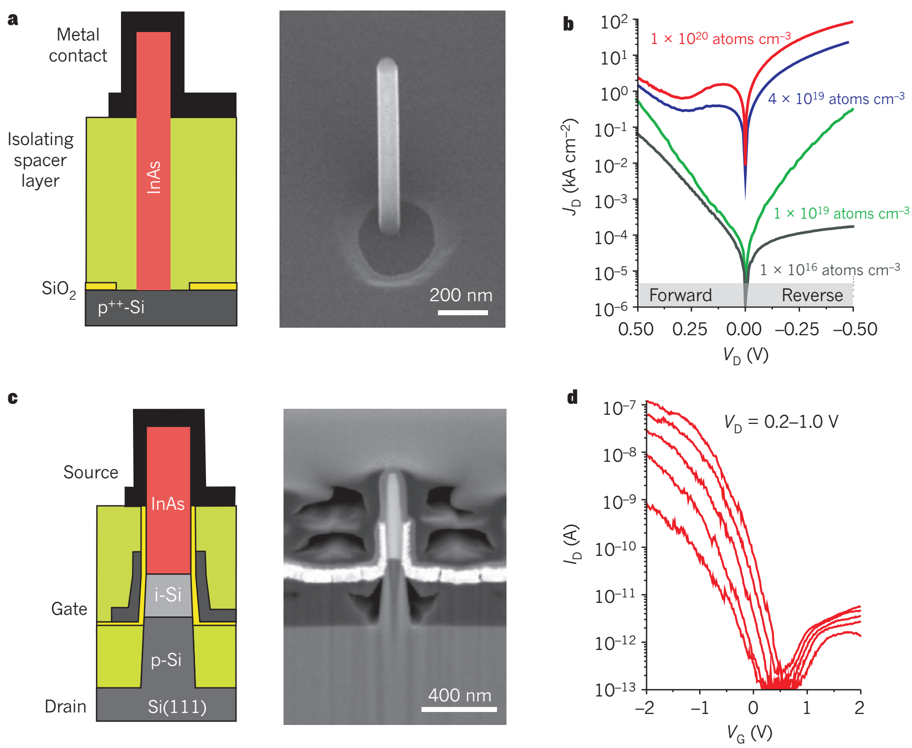


Figure 6 | InAs–Si heterojunction diodes and TFETs for improved performance. a, Left, schematic cross-section of an InAs–Si heterostructure nanowire diode. Right, scanning electron micrograph of an InAs nanowire grown epitaxially on Si. **b**, Current density–voltage (J_D–V) characteristics of Si–InAs heterojunction single-nanowire tunnel diodes. The high tunnel current densities required for TFETs are achieved for high doping levels. Different colours refer to different doping densities. **c**, Left, schematic cross-section of a vertical InAs–Si heterostructure nanowire TFET. Right, scanning electron micrograph showing a cross-section of the fabricated TFET. The InAs nanowire has a diameter of 100 nm, and the undoped Si channel on the p-type substrate is 150 nm long. **d**, Transfer characteristics, I_D–V_G, for various source–drain biases of the TFET shown in **c**.

evidence that BTBT is dominated by optical phonon-assisted inelastic transport, which can lead to a degradation of the S.

Furthermore, simulations have shown that the specific energy dependence of the one-dimensional density of states (which arises because the nanowire radius is so small) can be exploited to reach the quantum capacitance limit, resulting in better control of the channel potential. In this regime, the quantum capacitance, C_q, which is determined by the change of the charge in the channel resulting from a change in gate potential, is far smaller than the oxide capacitance, C_{ox}, due to the BTBT, and thus almost equals the total capacitance. This leads to a reduced total capacitance and improves the gate delay[72].

TFETs based on graphene nanoribbons theoretically offer the same benefits as carbon nanotube TFETs and may also provide planar processing compatibility. But so far, only theoretical studies have been performed, and these demonstrate the high potential of graphene nanoribbons in significantly improving I_{ON} to several hundred μA μm^{-1}, with an I_{OFF} of only a few pA μm^{-1} and a slope of less than 20 mV per decade[73].

However, for a practical implementation using graphene nanoribbons, the influence of the line edge roughness on the bandgap and the transport properties, and thus on the TFET performance, must be considered. Luisier and Klimeck[74] found that with rougher edges, the off current significantly increases because of a lowering of the graphene nanoribbon bandgap and an increase in the source-to-drain tunnelling leakage through the gate potential barrier. This leads to a deterioration of S and I_{ON}/I_{OFF}, which are no longer sufficient and thus limit the switching performance of graphene nanoribbon TFETs. The theoretical investigations done so far show that graphene nanoribbon TFETs have great potential, although for a practical implementation, significant technical challenges need to be met. Fiori and Iannaccone[75] suggested using bilayer graphene to fabricate TFETs, presenting an attractive alternative to graphene nanoribbons. The benefit of this approach is that the required energy gap is opened by applying a vertical electric field to the graphene bilayer, rather than by preparing small ribbons, which would require single-atom precision patterning.

Energy-efficient integrated circuits

The 65-nm CMOS transistor and the 50-nm C-TFET in both silicon and Ge/InAs can be compared directly by simulating the main characteristics of inverter cells (the basic building blocks of a circuit) operating at the same V_{DD}. Figure 7a shows the voltage transfer characteristics of inverters in these three implementations. The Ge/InAs C-TFET inverter has the most abrupt transition between the 1 and 0 states because it has a steep slope combined with a reasonably high I_{ON}, which results in the best noise margins and the highest dV_{OUT}/dV_{IN} gain. However, the transient response of the Ge/InAs C-TFET inverter is worse than that of the CMOS at a V_{DD} of 1 V (or even 0.5 V) with an associated delay of 358 ps. The advantages of TFET logic should be explored at lower frequencies corresponding to low-power and low-standby-power CMOS specifications[76], especially in terms of power savings.

A specific characteristic of TFETs that influences their transient response was reported by Mookerjea *et al.*[77], who were the first to observe that in a TFET the gate capacitance, C_{GG}, is dominated by the gate-to-drain capacitance, C_{GD}, under all bias conditions. This is in strong contrast to a MOSFET, where the C_{GD} and the gate-to-source capacitance, C_{GS}, have relatively balanced contributions to the gate capacitance. In a TFET, C_{GD} predominates even near the off state because the source-to-channel barrier resistance is large and the channel-to-drain barrier resistance is low. Therefore, the effective load capacitance for unloaded TFET inverters can be more than twice the gate capacitance because of the enhanced Miller effect (an increase in the equivalent input capacitance of an inverting voltage amplifier resulting from the amplification of capacitance between the input and output terminals) and the effective drive current. This can degrade the delay time of TFET inverters and generate current overshoots.

In general, there is an agreement in all published studies that TFETs are attractive for low-standby-power applications. Koswatta *et al.*[78] proposed a performance comparison between p-i-n TFETs with a carbon nanotube channel and conventional thermionic silicon MOSFETs. Their choice of a carbon nanotube channel is motivated by its direct energy bandgap and small carrier mass. They based their study on a comprehensive simulation framework and on experimental results. Specifically, they investigated the delay time, τ, and the switching energy calculated as the power-delay product. They found the intrinsic delay to be similar for both devices, but the TFET became much slower when load capacitance was considered. They also reported that phonon scattering degrades τ for both devices, but this increase is more important for TFETs. When operated under the quantum capacitance limit, TFETs have smaller switching energies than MOSFETs.

Until recently, investigations of circuit performance have been limited by the lack of availability of accurate compact models for TFETs calibrated on fabricated n- and p-type TFETs. One way to circumvent this problem was to build look-up table (LUT) models for current–voltage and capacitance–voltage characteristics using simulation data. One such study[6] used a LUT model for a type II heterojunction tunnel transistor (HETT) and compared its power reduction with a commercial bulk CMOS 45-nm technology in simulated ring oscillators. At $V_{DD} = 1$ V, a high-performance CMOS ring oscillator has a period of 450 ps and

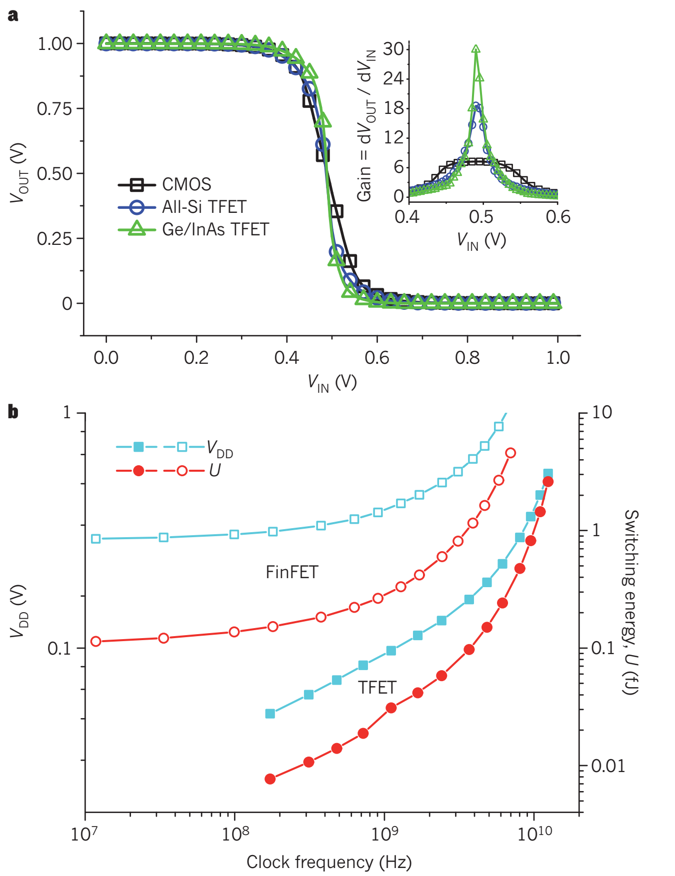


Figure 7 | Circuit-level characteristics of low-power TFETs. **a**, Comparison of the voltage transfer characteristics, V_{OUT}–V_{IN}, of CMOS and TFET inverters corresponding to the complementary device characteristics in Fig. 4b. The Ge/InAs C-TFET inverter has the most abrupt transition from the 1 to the 0 state with the highest differential gain, dV_{OUT}/dV_{IN} (inset) and best noise margins. **b**, Switching energy, U (red), and power supply voltage, V_{DD} (blue), against clock frequency simulated with a calibrated compact model for a heterostructure (In,Al)As/(Ga,Al)Sb TFET (filled symbols) and a silicon FinFET with a 20-nm channel length (open symbols); 16 cores of 1.5×10^6 circuits each are assumed. The total chip power was constrained, and V_{DD}, channel length, threshold voltage (V_T), Fermi energy level in the source (V_e), width and design were optimized to maximize the clock frequency[79]. Figure reprinted, with permission, from ref. 79.

53.9 μW dynamic power consumption, whereas the ring oscillator with a HETT consumes only 5.74 μW at 0.355 V to maintain the same period, achieving a 9.4-fold dynamic power reduction.

The most recent compact model for the TFET captures the essential physical features of the heterojunction TFET using the realistic case of (In,Al)As/(Ga,Al)Sb[79]. As well as the principal tunnelling mechanism, the effects of source degeneracy, back-injection from the drain, and direct source–drain tunnelling are included. Perhaps the most interesting idea of this study was to run the TFET model in an optimizer program that adjusts the device design parameters to achieve optimal chip-level performance when power and power density are constrained. The study compared 20-nm-channel-length TFET technology with FinFETs (multigate FET devices in which the gate is wrapped around a semiconducting channel shaped like a fin) in terms of switching energy and V_{DD} when the total chip power was constrained, and V_{DD} and the design parameters were varied to maximize the clock frequency (Fig. 7b). The calibrated simulations demonstrate the advantage of heterojunction TFETs over FinFETs in terms of both lower V_{DD} and switching energy.

In addition, TFETs retain their excellent switching characteristics even at high temperature[80–83] because the tunnelling mechanism makes them almost insensitive to temperature changes. Not only is the *S* invariable with temperature but so is the on current, which should increase only slightly owing to the decrease of the energy bandgap with temperature, as recently demonstrated experimentally[33,84].

Other attractive applications of TFETs are in analog integrated circuits[81] such as ultralow-power voltage-controlled oscillators and voltage references that have to deliver a well-defined output voltage, independently of supply voltage, temperature and process variations.

Finally, TFETs offer a solution for critical leakage power savings in static random access memory (SRAM). Six-transistor (6T)[85] and 4T[86] SRAM cell designs with CMOS and TFET technologies have been compared in terms of layout, performance and power on silicon platforms. A 700-fold improvement in leakage reduction over CMOS technology with a voltage supply of 0.3 V was demonstrated in the silicon TFET SRAM[85].

Conclusions

Today TFETs represent the most promising steep-slope switch candidate, having the potential to use a supply voltage significantly below 0.5 V and thereby offering significant power dissipation savings. Because of their low off currents, they are ideally suited for low-power and low-standby-power logic applications operating at moderate frequencies (several hundred MHz). Other promising applications of TFETs include ultralow-power specialized analog integrated circuits with improved temperature stability and low-power SRAM.

The biggest challenge is to achieve high performance (high I_{ON}) without degrading I_{OFF}, combined with an *S* of less than 60 mV per decade over more than four decades of drain current. This requires the additive combination of the many technology boosters specific to complementary heterostructure TFETs, which are available or under research on advanced SOI CMOS platforms.

Carbon materials such as carbon nanotubes and graphene are well suited for use in high-performance TFETs because of their ultrathin body thickness and their one-dimensional transport characteristics. However, enormous challenges exist for the experimental implementation of carbon TFETs with all of the process parameters under control. Heterostructure TFETs offer the best performance compromise for complementary logic through advanced band engineering, using Ge and InAs sources on silicon platforms for n- and p-type TFETs in ultrathin films or nanowires (with $I_{ON}/I_{OFF} > 10^7$, $I_{ON} = 100\,\mu A\,\mu m^{-1}$ and $V_{DD} < 0.5$ V). Such TFETs could offer opportunities for a hybrid CMOS C-TFET design, with TFETs as an add-on ultralow-power device option on advanced CMOS platforms. ■

1. Sakurai, T. Perspectives of low power VLSI's. *IEICE Trans. Electron* **E87-C,** 429–436 (IEICE, 2004).
2. Bernstein, K., Cavin, R. K., Porod, W., Seabaugh A. C. & Welser, J. Device and architectures outlook for beyond CMOS switches. *Proc. IEEE* **98,** 2169–2184 (2010).
3. Seabaugh, A. C. & Zhang, Q. Low voltage tunnel transistors for beyond CMOS logic. *Proc. IEEE* **98,** 2095–2110 (2010).
4. Sze, S. M. *Physics of Semiconductor Devices*, 1st edn (John Wiley, 1969).
5. Lundstrom, M. S. The MOSFET revisited: device physics and modeling at the nanoscale. *Proc. IEEE Int. SOI Conf.* 1–3 (IEEE, 2006).
6. Kim, D. *et al.* Heterojunction tunneling transistor (HETT)-based extremely low power applications. *Proc. Int. Symp. Low Power Electron. Design* 219–224 (IEEE/ACM, 2009).
7. Bhuwalka, K., Schultze, J. & Eisele, I. A simulation approach to optimize the electrical parameters of a vertical tunnel FET. *IEEE Trans. Electron Devices* **52,** 1541–1547 (2005).
8. Boucart, K. & Ionescu, A. M. Double-gate tunnel FET with high-κ gate dielectric. *IEEE Trans. Electron Devices* **54,** 1725–1733 (2007).
9. Kam, H., King-Liu, T.-J., Alon, E. & Horowitz, M. Circuit-level requirements for MOSFET-replacement devices. *Tech. Digest IEEE Int. Electron Devices Meet.* 1 (IEEE, 2008).
10. Hanson, S., Seok, M., Sylvester, D. & Blaauw, D. Nanometer device scaling in subthreshold logic and SRAM. *IEEE Trans. Electron Devices* **55,** 175–185 (2008).
11. Chang, L. *et al.* Practical strategies for power-efficient computing technologies. *Proc. IEEE* **98,** 215–236 (2010).
12. Nose, K. & Sakurai, T. Optimization of V_{DD} and V_{TH} for low-power and high-speed applications. *Proc. Asia S. Pacif. Design Automat. Conf.* 469–474 (ACM, 2000).
13. Gopalakrishnan, K., Griffin, P. B. & Plummer, J. D. I-MOS: a novel semiconductor device with subthreshold slope lower than *kT*/*q*. *Tech. Digest IEEE Int. Electron Devices Meet.* 289–292 (IEEE, 2002).
14. Zener, C. A theory of electrical breakdown of solid dielectrics. *Proc. R. Soc. Lond. A* **145,** 523–529 (1934).
15. Salahuddin, S. & Datta, S. Use of negative capacitance to provide voltage amplification for low power nanoscale devices. *Nano Lett.* **8,** 405–410 (2008).
16. Salvatore, G. A., Bouvet, D. & Ionescu, A. M. Demonstration of subthreshold swing smaller than 60mV/decade in Fe-FET with P(VDF-TrFE)/SiO_2 gate stack. *Tech. Digest IEEE Int. Electron Devices Meet.* 1–4 (IEEE, 2008).
17. Rusu, A., Salvatore, G. A., Jimenez, D. & Ionescu, A. M. Metal-ferroelectric-meta-oxide-semiconductor field effect transistor with sub-60mV/decade subthreshold swing and internal voltage amplification. *IEEE Int. Electron Devices Meet.* 16.3.1–16.3.4 (IEEE, 2010).
18. Abele, N. *et al.* Suspended-gate MOSFET: bringing new MEMS functionality into solid-state MOS transistor. *Tech. Digest IEEE Int. Electron Devices Meet.* 479–481 (IEEE, 2005).
19. Chen, F. *et al.* Integrated circuit design with NEM relays. *IEEE/ACM Int. Conf. Computer-Aided Design* 750–757 (IEEE, 2008).
20. Pott, V., Hei Kam, N. R., Jaeseok, J., Alon, E. & Tsu-Jae, K. L. Mechanical computing redux: relays for integrated circuit applications. *Proc. IEEE* **98,** 2076–2094 (2010).
21. Quinn, J., Kawamoto, G. & McCombe, B. Subband spectroscopy by surface channel tunneling. *Surf. Sci.* **73,** 190–196 (1978).
22. Banerjee, S., Richardson W., Coleman J. & Chatterjee, A. A new three-terminal tunnel device. *IEEE Electron Device Lett.* **8,** 347–349 (1987).
23. Takeda, E., Matsuoka, H., Igura, Y. & Asai, S. A band to band tunneling MOS device B2T-MOSFET. *Tech. Digest IEEE Int. Electron Devices Meet.* 402–405 (IEEE, 1988).
24. Baba, T. Proposal for surface tunnel transistors. *Jpn. J. Appl. Phys.* **31,** L455–L457 (1992).
25. Reddick, W. & Amaratunga, G. Silicon surface tunnel transistor. *Appl. Phys. Lett.* **67,** 494–496 (1995).
26. Koga, J. & Toriumi, A. Negative differential conductance in three-terminal silicon tunneling device. *Appl. Phys. Lett.* **69,** 1435–1437 (1996).
27. Hansch, W., Fink, C., Schulze, J. & Eisele, I. A vertical MOS-gated Esaki tunneling transistor in silicon. *Thin Solid Films* **369,** 387–389 (2000).
28. Aydin, C. *et al.* Lateral interband tunneling transistor in silicon-on-insulator. *Appl. Phys. Lett.* **84,** 1780–1782 (2004).
29. Appenzeller, J., Lin, Y.-M., Knoch J. & Avouris, P. Band-to-band tunneling in carbon nanotube field-effect transistors. *Phys. Rev. Lett.* **93,** 196805 (2004).
30. Krishnamohan, T., Kim, D., Raghunathan, S. & Saraswat, K. Double-gate strained-Ge heterostructure tunneling FET (TFET) with record high drive currents and <60 mV/dec subthreshold slope. *Tech. Digest IEEE Int. Electron Devices Meet.* 947–949 (IEEE, 2008).
31. Mayer, F. *et al.* Impact of SOI, $Si_{1-x}Ge_x$OI and GeOI substrates on CMOS compatible tunnel FET performance. *Tech. Digest IEEE Int. Electron Devices Meet.* 163–166 (IEEE, 2008).
32. Hu, C. *et al.* Prospect of tunneling green transistor for 0.1 V CMOS. *IEEE Int. Electron Devices Meet.* 16.1.1–16.1.4 (IEEE, 2010).
33. Moselund, K. E. *et al.* Comparison of VLS grown Si NW tunnel FETs with different gate stacks. *Proc. Eur. Solid State Device Res. Conf.* 448–451 (IEEE, 2009).
34. Wang, P. F. *et al.* Complementary tunneling transistor for low power application. *Solid-State Electron.* **48,** 2281–2286 (2004).
35. Knoch, J. & Appenzeller, J. A novel concept for field-effect transistors – the tunneling carbon nanotube FET. *Digest Device Res. Conf.* 153–156 (IEEE, 2006).
36. Knoch, J., Mantl, S. & Appenzeller, J. Impact of the dimensionality on the performance of tunneling FETs: bulk versus one-dimensional devices. *Solid-State Electron.* **51,** 572–578 (2007).
37. Zhang, Q., Zhao, W. & Seabaugh, A. Low-subthreshold-swing tunnel transistors. *IEEE Electron Device Lett.* **27,** 297–300 (2006).
38. Luisier, M. & Klimeck, G. Simulation of nanowire tunneling transistors: from

the Wentzel–Kramers–Brillouin approximation to full-band phonon-assisted tunneling. *J. Appl. Phys.* **107,** 084507 (2010).
39. Appenzeller, J., Knoch, J., Björk, M. T., Riel, H. & Riess, W. Toward nanowire electronics. *IEEE Trans. Electron Devices* **55,** 2827–2845 (2008).
40. Ionescu, A. M., Boucart, K., Moselund, K. E. & Pott, V. *Small Swing Switches* (Cambridge Univ. Press, in the press).
41. Leonelli, D. *et al.* Optimization of tunnel FETs: impact of gate oxide thickness, implantation and annealing conditions. *Proc. Eur. Solid State Device Res. Conf.* 170–173 (IEEE, 2010).
42. Boucart, K. & Ionescu, A. M. Length scaling of the double gate tunnel FET with a high-*κ* gate dielectric. *Solid State Electron.* **51,** 1500–1507 (2007).
43. Sandow, C., Knoch, J., Urban, C., Zhao, Q.-T. & Mantl, S. Impact of electrostatics and doping concentration on the performance of silicon tunnel field-effect transistors. *Solid State Electron.* **53,** 1126–1129 (2009).
44. Bhuwalka, K., Schulze, J. & Eisele, I. Performance enhancement of vertical tunnel field-effect transistor with SiGe in the dp^+ layer. *Jpn. J. Appl. Phys.* **43,** 4073–4078 (2004).
45. Verhulst, A. *et al.* Complementary silicon-based heterostructure tunnel-FETs with high tunnel rates. *IEEE Electron Device Lett.* **29,** 1398–1401 (2008).
46. Knoch, J. Optimizing tunnel FET performance–impact of device structure, transistor dimensions and choice of material. *Int. Symp. VLSI-TSA* 45–46 (IEEE, 2009).
47. Knoch, J. & Appenzeller, J. Modeling of high-performance p-type III–V heterojunction tunnel FETs. *IEEE Electron Device Lett.* **31,** 305–307 (2010).
48. Koswatta, S. O., Koester, S. J. & Haensch, W. On the possibility of obtaining MOSFET-like performance and sub-60-mV/dec swing in 1-D broken-gap tunnel transistors. *IEEE Trans. Electron Devices* **57,** 3222–3223 (2010).
49. Hu, C. Green transistor as a solution to the IC power crisis. *Proc. 9th Int. Conf. Solid-State Integrated-Circuit Technol.* 16–20 (IEEE, 2008).
50. Hu, C. *et al.* Prospect of tunneling green transistor for 0.1 V CMOS. *IEEE Int. Electron Devices Meet.* 16.1.1–16.1.4 (IEEE, 2010).
51. Asra, R. *et al.* A tunnel FET for V_{DD} scaling below 0.6V with a CMOS-comparable performance. *IEEE Trans. Electron Devices* **58,** 1855–1863 (2011).
52. De Michielis, L., Lattanzio, L., Palestri, P., Selmi L. & Ionescu, A. M. Tunnel-FET architecture with improved performance due to enhanced gate modulation of the tunneling barrier. *IEEE Device Res. Conf.* (IEEE, in the press).
53. Nayfeh, O. M. *et al.* Design of tunneling field-effect transistors using strained-silicon/strained-germanium type-II staggered heterojunctions. *IEEE Electron Device Lett.* **29,** 1074–1077 (2008).
54. Boucart, K., Ionescu, A. M. & Riess, W. Asymmetrically strained all-silicon tunnel FETs featuring 1V operation. *Proc. Eur. Solid State Device Res. Conf.* 452–456 (IEEE, 2009).
55. Boucart, K., Riess, W. & Ionescu, A. M. Lateral strain profile as key technology booster for all-silicon tunnel FETs. *IEEE Electron Device Lett.* **30,** 656–658 (2009).
56. Boucart, K. Simulation of a Double Gate Silicon Tunnel FET with a High-*κ* Dielectric. PhD thesis, Ecole Polytechnoique Fédérale de Lausanne (2009).
57. Le Royer, C. & Mayer, F. Exhaustive experimental study of tunnel field effect transistors (TFETs): from materials to architecture. *Proc. 10th Int. Conf. Ultimate Integration Silicon* 53–56 (IEEE, 2009).
58. Loh, W.-Y. *et al.* Sub-60nm Si tunnel field effect transistors with $I_{on} > 100$ μA/ μm. *Proc. Eur. Solid State Device Res. Conf.* 162–165 (IEEE, 2010).
59. Mookerjea, S. *et al.* Experimental demonstration of 100 nm channel length $In_{0.53}Ga_{0.47}As$-based vertical inter-band tunnel field effect transistors (TFET) for ultra low-power logic and SRMA applications. *IEEE Int. Electron Devices Meet.* 137.1–137.4 (IEEE, 2009).
60. Zhao, H. *et al.* InGaAs tunneling field-effect transistors with atomic-layer-deposited gate oxides. *IEEE Trans. Electron Devices* **58,** 2990–2995 (2011).
61. Mookerjea, S., Mohata, D., Mayer, T., Narayanan V. & Datta, S. Temperature-dependent characteristics of a vertical tunnel FET. *IEEE Electron Device Lett.* **31,** 564–566 (2010).
62. Wang, L., Yu, E., Taur Y. & Asbeck, P. Design of tunneling field-effect transistors based on staggered heterojunctions for ultralow-power applications. *IEEE Electron Device Lett.* **31,** 431–433 (2010).
63. Mohata, D. *et al.* Experimental staggered-source and N^+ pocket-doped channel III–V tunnel field-effect transistors and their scalabilities. *Appl. Phys. Express* **4,** 024105 (2011).
64. Zhou, G. *et al.* Self-aligned $InAs/Al_{0.45}Ga_{0.55}Sb$ vertical tunnel FETs. *IEEE Device Res. Conf.* 205–206 (IEEE, 2011).
65. Tomioka, K., Motohisa, J., Hara S. & Fukui, T. Control of InAs nanowire growth directions on Si. *Nano Lett.* **8,** 3475–3480 (2008).
66. Björk, M. T. *et al.* Si–InAs heterojunction Esaki tunnel diodes with high current densities. *Appl. Phys. Lett.* **97,** 163501 (2010).
67. Bessire, C. D. *et al.* Trap-assisted tunneling in Si–InAs nanowire heterojunction tunnel diodes. *Nano Lett.* **11,** 4195–4199 (2011).
68. Lu, Y. *et al.* Geometry dependent tunnel FET performance — dilemma of electrostatics *vs*. quantum confinement. *IEEE Device Res. Conf.* 17–18 (IEEE, 2010).
69. Schmid, H. *et al.* Fabrication of vertical InAs–Si heterojunction tunnel field effect transistors. *IEEE Proc. Device Res. Conf.* 181–182 (2011).
70. Poli, S. *et al.* Computational study of the ultimate scaling limits of CNT tunneling devices. *IEEE Trans. Electron Devices* **55,** 313–321 (2008).
71. Koswatta, S. O., Lundstrom, M. S. & Nikonov, D. E. Band-to-band tunneling in a carbon nanotube metal-oxide-semiconductor field-effect transistor is dominated by phonon-assisted tunneling. *Nano Lett.* **7,** 1160–1164 (2007).
72. Appenzeller, J., Lin, Y.-M., Knoch, J., Chen, Z. & Avouris, P. Comparing carbon nanotube transistors — the ideal choice: a novel tunneling device design. *IEEE Trans. Electron Devices* **52,** 2568– 2576 (2005).
73. Zhang, Y. *et al.* Giant phonon-induced conductance in scanning tunneling spectroscopy of gate-tunable graphene. *Nature Phys.* **4,** 627–630 (2008).
74. Luisier, M. & Klimeck, G. Performance limitations of graphene nano ribbon tunneling FETS due to line edge roughness. *IEEE Device Res. Conf.* 201–202 (IEEE, 2009).
75. Fiori, G. & Iannaccone, G. Ultralow-voltage bilayer graphene tunnel FET. *IEEE Electron Device Lett.* **30,** 1096–1098 (2009).
76. ITRS International Technology Working Groups. *International Technology Roadmap for Semiconductors* ⟨http://www.itrs.net⟩ (2010).
77. Mookerjea, S., Krishnan, R., Datta, S. & Narayanan, V. On enhanced Miller capacitance effect in interband tunnel transistors. *IEEE Electron Device Lett.* **30,** 1102–1104 (2009).
78. Koswatta, S., Lundstrom, M. & Nikonov, D. Performance comparison between p-i-n tunneling transistors and conventional MOSFETs. *IEEE Trans. Electron Devices* **56,** 456–465 (2009).
79. Solomon, P. M., Frank, D. J. & Koswatta, S. O. Compact model and performance estimation for tunneling nanowire FET. *IEEE Device Res. Conf.* 197–198 (IEEE, 2011).
80. Born, M. *et al.* Tunnel FET: a CMOS device for high temperature applications. *Proc. 25th Int. Conf. Microelectron.* 124–127 (IEEE, 2006).
81. Fulde, M. *et al.* Fabrication, optimization, and application of complementary multiple-gate tunneling FETs. *Proc. INEC* 579–584 (IEEE, 2008).
82. Kane, E. O. Zener tunneling in semiconductors. *J. Phys. Chem. Solids* **12,** 181–188 (1959).
83. Mookerjea, S., Mohata, D., Mayer, T., Narayanan, V. & Datta, S. Temperature-dependent I–V characteristics of a vertical $In_{0.53}Ga_{0.47}As$ tunnel FET. *IEEE Electron Device Lett.* **31,** 564–566 (2010).
84. Moselund, K. E. *et al.* Silicon nanowire tunnel FETs: low-temperature operation and influence of high-*k* gate dielectric. *IEEE Trans. Electron Devices* **58,** 2911–2916 (2011).
85. Singh, J. *et al.* A novel Si-tunnel FET based SRAM design for ultra low-power 0.3 V V_{DD} applications. *Proc. Asia S. Pacif. Design Automat. Conf.* 181–186 (ACM, 2010).
86. Saripalli, V., Mohata, D. K., Mookerjea, S., Datta, S. & Narayanan, V. Low power loadless 4T SRAM cell based on degenerately doped source (DDS) $In_{0.53}Ga_{0.47}As$ tunnel FETs. *IEEE Device Res. Conf.* 101–102 (IEEE, 2010).

Acknowledgements Some of this work was supported by the European Commission under the FP7 projects *Guardian Angels for a Smarter Life* and *STEEPER*. K. Boucart, L. De Michielis, C. Le Royer, K. Moselund, M. Björk, H. Schmid, W. Riess and P. Solomon are particularly acknowledged for useful discussions and supporting materials.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to A.M.I. (adrian.ionescu@epfl.ch).

doi:10.1038/nature10680

A role for graphene in silicon-based semiconductor devices

Kinam Kim[1], Jae-Young Choi[1], Taek Kim[1], Seong-Ho Cho[1] & Hyun-Jong Chung[1]

As silicon-based electronics approach the limit of improvements to performance and capacity through dimensional scaling, attention in the semiconductor field has turned to graphene, a single layer of carbon atoms arranged in a honeycomb lattice. Its high mobility of charge carriers (electrons and holes) could lead to its use in the next generation of high-performance devices. Graphene is unlikely to replace silicon completely, however, because of the poor on/off current ratio resulting from its zero bandgap. But it could be used to improve silicon-based devices, in particular in high-speed electronics and optical modulators.

Silicon-based microprocessors and memory chips with a linewidth as small as 20 nm can meet the demand for low-power multifunctional chips that can process and store massive amounts of heterogeneous data. But achieving physical dimensions near the 'deeper nanoscale' regime of 10 nm or beyond[1] is a challenge for existing technologies. It will require dimensional scaling using novel structures, materials and processes for complementary metal–oxide–semiconductor (CMOS) devices, such as three-dimensional architectures for memory and logic, high-κ/metal gate transistors, extreme ultraviolet lithography and new computing architectures. Materials such as graphene, which has an extremely high charge-carrier mobility, are expected to have an important role in the advancement of semiconductor technology[2–4].

Graphene is a monolayer of carbon atoms arranged in a two-dimensional honeycomb lattice and is a basic building block of well-known carbon materials such as graphite, carbon nanotubes and fullerene. Since graphene was isolated by mechanical exfoliation in 2004 (ref. 5), many extraordinary properties have been reported, such as extremely high electron mobility[3–5]. High mobility arises because electrons can propagate without scattering over large distances, perhaps micrometres, because of its reduced phonon scattering. A recent poll conducted by the semiconductor industry for the *International Technology Roadmap for Semiconductors* (ITRS) named graphene as the material likely to have the greatest impact on geometric scaling, thanks to its high mobility, which is desirable in metal–oxide–semiconductor field-effect transistor (MOSFET) channels[6]. Furthermore, graphene's strong interactions with photons[7–9] and electrochemical stability could add more functions to silicon-based CMOS devices, such as radio-frequency switches and photonic modulators.

Here we consider how graphene can be incorporated into these semiconductor devices, examine the requirements and challenges that must be met, and discuss possible solutions.

Radio-frequency transistors

The high mobility of charge carriers in graphene is ideal for obtaining fast switching and a high 'on' current (I_{ON}). Zero bandgap induces a large 'off' current, however, so the on/off current ratios for graphene transistors are about 100 (ref. 5), much lower than the 10^3–10^6 required for mainstream logic applications. But this relatively low on/off ratio is not a significant problem for high-frequency applications, such as radio-frequency switches.

In general, radio-frequency transistor performance is characterized by two parameters: the cutoff frequency (f_T) and the maximum oscillation frequency (f_{max}), where f_T and f_{max} represent how fast channel current and power transmission, respectively, are modulated by the gate. As shown in following equation

$$f_T = \frac{g_m}{2\pi C_G} \quad (1)$$

f_T is determined by the gate capacitance (C_G) and the transconductance (g_m). Thus far, the highest measured f_T has been 300 GHz with Co_2Si-nanowire gates using exfoliated graphene[10]. At a wafer scale, 240 GHz has been reported using epitaxial graphene[11], compared with 200 GHz using chemical vapour deposition (CVD)[12]. As shown in Table 1, even with larger gate lengths, graphene has demonstrated superior performance to silicon[13] and III–V group compounds[14–16] as a result of its high drift velocity of 4×10^7 cm s^{-1} (ref. 17). This means that f_T for graphene transistors is approximately 1.42 THz for a 56-nm gate[17], although this value excludes parasitic capacitances, which would be halved in such cases.

Another important parameter for radio-frequency transistors is f_{max}:

$$f_{max} = \frac{f_T}{2\sqrt{(g_D(R_G + R_{SD}) + 2\pi f_T R_G C_G)}} \quad (2)$$

where g_D is the channel conductance, R_G is the gate resistance and R_{SD} is the source–drain resistance. Studies have shown large discrepancies in values between f_T and f_{max} for graphene radio-frequency transistors. Previous work initially attributed the discrepancy to the high contact resistance[18] and high R_G without considering issues inherent to the graphene film, in particular the low output resistance. We expect that the low output resistance is one of the key factors in increasing f_{max}. Because f_{max} measures transmitted power (I^2R), it can be increased by minimizing g_D attained at channel saturation; this is a pinch-off in most MOS transistors. In graphene, the pinch-off condition, effective for the drain saturation, can be produced if graphene has a bandgap. More research to improve f_{max} through improvements in saturation currents is needed[19,20]. As long as f_{max} remains at current levels, it will be difficult to use graphene transistors in radio-frequency amplifiers. At present, they will be valuable for radio-frequency switching applications only if f_T exceeds the performance of current silicon-based devices.

[1]Samsung Advanced Institute of Technology (SAIT), Samsung Electronics, Yongin-Si, Gyeonggi-Do 446-712, South Korea.

Table 1 | Properties of radio-frequency transistors

Transistor	f_T (GHz)	f_{max} (GHz)	L_g (nm)	W_g (μm)	Fingers	g_m (mS μm^{-1})	Mobility (cm^2 V^{-1} s^{-1})	Specific contact resistance (Ω μm^2)
Graphene								
Estimated[17]*	1,420	–	56	2	1	2.3	>10,000	
Measured[12]*	300	–	144	10	1	1.27		7.5 (ref. 23)
CVD grown[18]	155	<10	40	30	2	0.02	500–600	
Epitaxial[21]	100	–	240	–	2	0.15	1,000–1,500	
Silicon								
Silicon[13]	485	–	29	30	30	1.3	1,400‡	0.1§
ITRS 2011	310	330	29	–	–	–		
ITRS 2014†	480	540	18	–	–	–		
III–V								
InP[14]	385	>1,100	<50	40	2	1.2	15,000	0.5 (ref. 16)
InAs[15]	628	331	30	100	2	1.62	13,200	

*Flake graphene; †Target specification; ‡Electron mobility; §100 nm × 100 nm NiSi/Si. L_g, gate length; W_g, gate width.

As shown in Table 1, the estimated f_T for graphene is significantly higher than that for silicon. However, the experimentally measured f_T for graphene is much lower than expected. To increase f_T, mobility has to be increased[11,12,21], for which three factors need to be considered: minimizing defects in graphene; minimizing impurities that cause scattering at the graphene channel–gate dielectric interface[22]; and reducing contact resistance[23,24].

Defects in graphene are highly dependent on the growth process. The highest reported mobility is 200,000 cm^2 V^{-1} s^{-1}, which has been obtained using mechanically exfoliated graphene flakes without a substrate at a temperature of 5 K (ref. 6). Mobility remains within the same order of magnitude at elevated temperatures because the scattering mechanism is mainly caused by electron or hole puddles within the two-dimensional graphene sheet[25]. In reality, graphene flakes cannot be used for silicon applications, so CVD or sublimation of SiC will be needed, even though these might reduce mobility.

The impurities that cause scattering at the graphene channel–gate dielectric interface have been attributed to interactions with the underlying substrate, such as surface charge traps[26,27], interfacial phonons[28] and nanometre-scale deformation or ripples[29–31]. To minimize these impurities, we propose a new structure containing air gaps, a 'nothing-on-graphene' architecture that provides a free-standing condition for graphene. On the other side of graphene is a gate oxide that does not reduce the mobility, such as hexagonal boron nitride (hBN), which allows a reasonable charge transport in graphene (up to 40,000 cm^2 V^{-1} s^{-1} at room temperature) because it is atomically flat and free of dangling bonds and charge traps[32,33].

Figure 1 shows the formation of the 'nothing-on-graphene' architecture. Copper gates are formed through a 'damascene' process, along with the first metal line (M1), which here is copper (Fig. 1a). The next step is the transfer of hBN and graphene (Fig. 1b). The critical step in producing 'nothing-on-graphene' is the formation of air gaps between the source and drain metal electrodes; this is achieved by depositing a sacrificial film on the electrodes (Fig. 1c), which is then etched back to expose the electrodes (Fig. 1d). A porous film is then deposited (Fig. 1d) over the electrodes, and the sacrificial film is removed, creating the air gaps (Fig. 1e). The porous film must then be patterned (Fig. 1f). This 'nothing-on-graphene' architecture minimizes interfacial scattering interactions and thus maximizes mobility.

The contact resistance of graphene is more than 70 times higher than that of silicon (see Table 1). However, reducing the contact resistance poses significant challenges because of the interactions between graphene and different metals. Graphene shows weak binding with metals such as copper, gold and platinum on their (111) surfaces, where

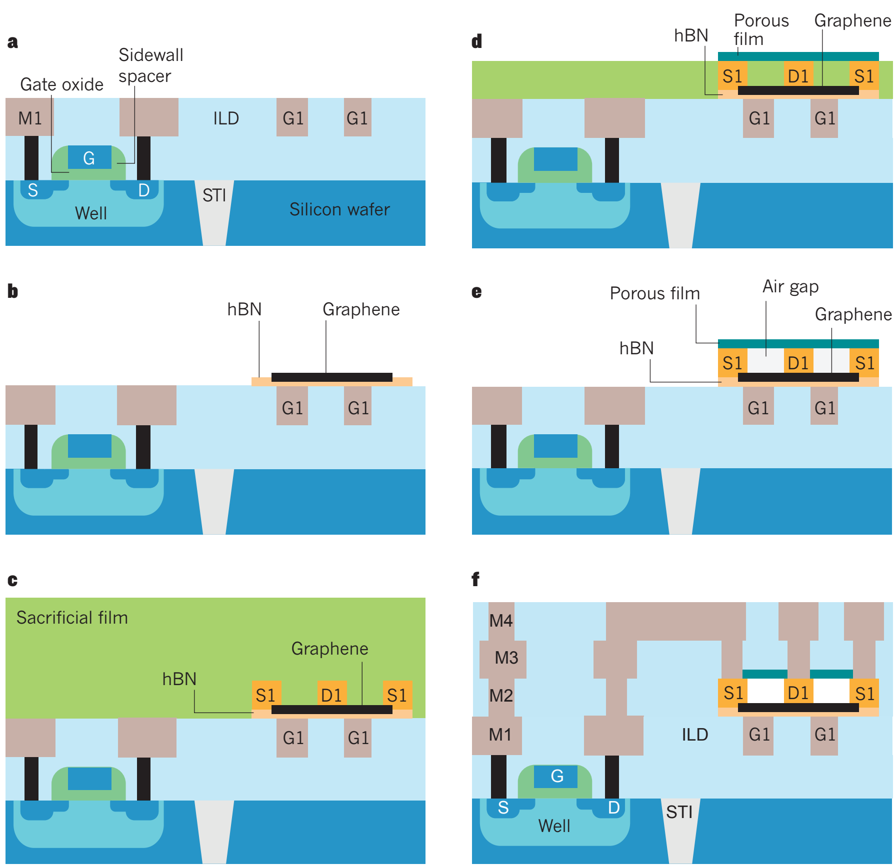


Figure 1 | Formation of 'nothing-on-graphene' architecture, which minimizes the scattering of electrons. **a**, A 'damascene' process is used to create a layer of copper metal (M1) that functions as the gate electrode (G1) for the graphene transistor. G, S and D indicate the gate, the source and the drain of the silicon transistor, respectively, and STI indicates the shallow trench isolation. The well is for the separation of the p–n doping regions. **b**, A layer of hexagonal boron nitride (hBN) is introduced to act as a gate oxide. This is atomically flat and lacks dangling bonds and charge traps, so it allows fast charge transport. A layer of graphene is placed over the hBN. **c**, The source (S1) and drain (D1) metal electrodes are patterned, and a sacrificial film is placed over the top. **d**, The sacrificial film is then etched back to reveal the electrodes, and a porous film is then deposited. **e**, The remainder of the sacrificial film is removed through pores of the porous film, leaving air gaps between the electrodes. **f**, Once the 'nothing-on-graphene' architecture is established, CMOS 'back end of the line' processes, including the addition of further metal layers, can complete the circuit. ILD, interlayer dielectric.

Table 2 | On/Offcurrent ratios for a 20-nm Si logic process

Property	Type 1*	Type 2†	Type 3‡	Type 4§
On/Off ratio	$\sim 2 \times 10^6$	$\sim 2 \times 10^5$	$\sim 4 \times 10^4$	$\sim 5 \times 10^3$
I_{ON} ($\mu A\ \mu m^{-1}$)	~800	~1,000	~1,100	~1,300

All transistor types are needed for 20-nm silicon CMOS logic circuits. *Low on current, highest on/off ratio; †Medium on current, high on/off ratio; ‡Medium on current, medium on/off ratio; §High on current, low on/off ratio.

graphene preserves its band structure[34,35]. By contrast, graphene binds strongly to nickel, cobalt and palladium as their electronic structures are strongly intermixed[36,37]. Although contact resistance is expected to be lower in strongly binding metals, binding strength has been shown not to affect resistance, as experimental values for gold, nickel and palladium are quite similar[24,36–39]. This suggests that even for strong metal–graphene contacts, electron transmission from the graphene region in contact with the metal to the pristine graphene channel becomes difficult. Careful studies are required to determine whether contact resistances have intrinsic limitations.

Even if f_T is maximized, the task of improving f_{max} still remains. To accomplish this, saturation (a pinch-off condition with a bandgap) should be created in graphene for a high output resistance.

When a bandgap is effectively formed, graphene can be used in high-speed logic applications. So far, there have been no specific guidelines on how wide the bandgap should be, but this could be inferred from the minimum I_{ON}/I_{OFF} ratio requirement. For a noticeable bandgap to be engineered at the graphene–metal junction, I_{OFF} must be governed by the thermionic emission of carriers through the metal–graphene Schottky barrier. Thus I_{OFF} would be proportional to $\exp(-q\varphi_{barrier}/kT)$, where q is the electron charge, $\varphi_{barrier}$ is the Schottky barrier height, k is the Boltzmann constant and T is the temperature. In addition, assuming that the work functions of graphene and the metal are the same, the resultant Schottky barrier height would be $E_g/2$, where E_g is the bandgap energy. Hence the I_{ON}/I_{OFF} ratio would be proportional to $\exp(E_g/2kT)$. Among various transistor technologies of 20-nm CMOS logic products, a high-speed transistor requires the I_{ON}/I_{OFF} ratio to be three to four orders of magnitude, which is the lowest acceptable value for high-performance logic applications (K. Kim, unpublished data; see Table 2). Because the I_{ON}/I_{OFF} ratio is proportional to $\exp(E_g/2kT)$, the minimum required bandgap would be 360 meV for high-speed CMOS applications.

There is considerable research to improve the I_{ON}/I_{OFF} ratios of current graphene transistors: nanoribbon[40–42], bilayer[43] and perforated graphene[44–46]. So far, such studies have been unable to achieve this value. It might take years for graphene transistors to be used for high-speed logic, but they could still find use embedded in radio-frequency applications such as low-noise amplifiers, mixers, frequency multipliers and resonators[47–49].

Once a graphene radio-frequency transistor can be integrated into a silicon CMOS, and assuming it has a very high f_T and linear gain, which are usually observed in graphene radio-frequency transistors, it will enhance the performance of low-noise and low-distortion circuits. Furthermore, if f_{max} can also be improved, its usage will extend to all circuitry, including all power-sensitive blocks.

Optical devices for silicon photonics

As well as the radio-frequency field, graphene could find another promising market in photonics because it can facilitate ultrawide bandwidths for exa-scale (10^{18}) computing systems. The explosive data growth on the Internet drives the need for exa-scale systems for large data centres. Peta-scale (10^{15}) systems consumed about 2.5 MW in 2008; a simple estimate shows that exa-scale systems will consume about 2.5 GW, which is unacceptably high[50]. Energy efficiency is therefore the next challenge for the silicon industry. Silicon photonics can provide ultrawide bandwidths through the multiplexing of numerous wavelengths[51,52] and reduced power consumption.

Graphene can find a use here because of its optical properties. It absorbs about 2.3% of normal incident light, despite only being one atom thick[53–55]. Transparency is almost independent of wavelength over spectra ranging from visible light to infrared because there are two-dimensional gases of free electrons in the gapless electronic structure of graphene[53,56]. These optical characteristics, allied with its high electron mobility, make graphene a prime candidate for ultrawide-bandwidth optical modulators and photo-detectors.

An optical modulator is one of the key components for altering the properties of light, such as its phase, amplitude or polarization, by electro-refraction or electro-absorption[57]. Optical modulators, such as Mach–Zehnder interferometers[57,58], ring or disk resonators[59,60] and germanium- or III–V-based electro-absorption modulators[61,62], are based on interference, resonance and bandgap absorption, respectively. Their operating spectra are usually narrow[59–63]. A single sheet of graphene on a silicon waveguide has recently been reported to provide an ultrafast response time[7].

The interband transitions of photo-generated electrons are modulated over broad spectral ranges by a drive voltage, so a broadband and high-speed optical modulator can be implemented with graphene. This CMOS-compatible graphene optical modulator can provide excellent performance over broad operating spectra ranging from 1.35 μm to 1.60 μm, with an operating speed of 1.2 GHz, a modulation efficiency of ~0.1 dB μm^{-1} and a small footprint of 25 μm^2 (ref. 7).

Optical modulators can be further improved by increasing both modulation depth and operating speed using a novel modulator structure, as shown in Fig. 2, in which graphene is placed at the location of maximum light intensity of the ridge region in the waveguide,

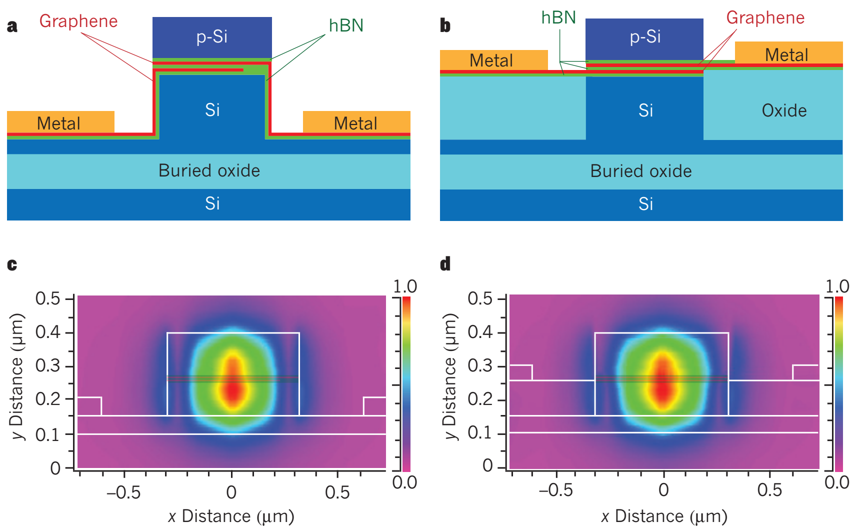


Figure 2 | Structure of a graphene-gated optical modulator. **a**, A ridge-type modulator. **b**, A buried-type modulator. In each case, two monolayer graphene sheets and an hBN spacer 7 nm thick are used to separate the waveguides, which are made of two different sections of silicon substrate and polycrystalline silicon (p-Si). The ridge is 600 nm wide, 250 nm high and 35 μm long. The metal electrode is located 300 nm from the waveguide. **c**, **d**, Transverse electric mode profiles at a wavelength of 1.55 μm are shown for the ridge type (**c**) and the buried type (**d**) modulator. The multilayer structure of graphene and hBN is overlaid with the mode profiles.

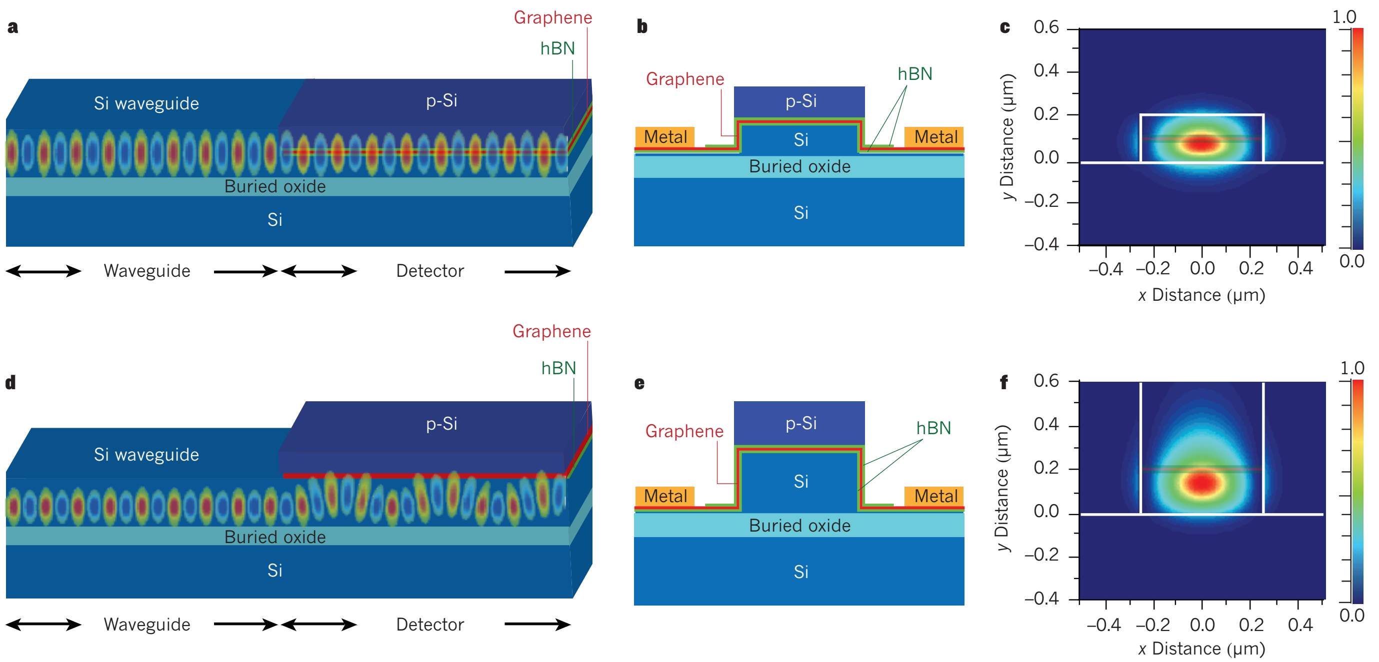


Figure 3 | The structure of graphene photo-detectors with integrated waveguides. **a–c**, Butt-coupled photo-detectors. **a**, A longitudinal schematic is shown with an overlaid longitudinal electric field (E-field) that shows the light transport from the waveguide to the detector. The layer of graphene is surrounded by hBN to maximize mobility. **b**, A cross-section of the photo-detector shown in **a**. Graphene is placed between the half-etched Si waveguide and the p-Si to minimize optical losses at the waveguide–detector interface. **c**, A cross-sectional E-field profile showing a good overlap with the graphene. **d–f**, Evanescent-coupled photo-detectors. **d**, A longitudinal schematic is shown with an overlaid longitudinal E-field profile. The E-field extends into the p-Si, enhancing the interaction with graphene. **e**, A cross-section of the photo-detector shown in **d**. Unlike **b**, graphene is placed above the Si waveguide. **f**, A cross-sectional profile showing the E-field tailing upwards to increase the overlap with the graphene. For simplicity, longitudinal views omit the metal grids. The silicon waveguide is 500 nm wide and 200 nm high. The E-field amplitudes are calculated at a wavelength of 1.3 μm. Graphene's thickness of 0.335 nm and absorption coefficient of 301,655 cm^{-1} (ref. 56) are used to calculate the normalized detector lengths. The multilayer structure of graphene and hBN is overlaid on the cross-sectional E-field profiles.

maximizing the modulation depth. To this end, the ridge is composed of two different regions, for example polycrystalline silicon on the top and single-crystalline silicon on the bottom. New modulator structures where dual graphene layers are separated by a thin hBN spacer are shown in Fig. 2a, b. Their propagating mode profiles, which confirm a good spatial overlap between the light mode and the graphene layers, are shown in Fig. 2c,d. Oxide deteriorates carrier mobility in graphene, so hBN is used as an insulator because it allows graphene to maintain its high mobility. The capacitance-resistance (*RC*) time constant can be reduced by replacing high-resistance bulk silicon with low-resistance graphene, and by replacing the gating spacer with low-dielectric-constant hBN. With two single sheets of graphene 600 nm wide on the waveguide, which is 300 nm from the metal contact and 35 μm long (Fig. 2b), the total resistance, including the metal contact, is reduced from 600 Ω to 26 Ω, where the sheet resistance is 15 Ω and the contact resistance is 11 Ω, using unit sheet and unit contact resistances for single-sheet graphene of 300 Ω per square ($\Omega\,\square^{-1}$) (ref. 64) and 200 Ω μm (ref. 36), respectively. In addition, the capacitance is reduced from 220 fF to 100 fF. The predicted bandwidth of this high-speed modulator, $f_{3B} = 1/2\pi RC$, is intrinsically suitable for 55-GHz modulation, which is comparable to the best reported optical modulator bandwidth[63]. In the near future, it is likely that both inter-chip and intra-chip interconnects for ultrawide-bandwidth data networks will be possible using graphene.

Another prime optical application for graphene is as a photo-detector. Extremely high bandwidth is possible because the high carrier mobility allows the ultrafast extraction of light-generated carriers. The transit time-limited bandwidth is calculated to be 1.5 THz[4] at the reported saturation carrier velocity[65]. In comparison, the best results using germanium photo-detectors are around 30 GHz[66,67], and the maximum bandwidth is expected to be 80 GHz[68].

The feasibility of graphene as a photo-detector has been demonstrated in photo-generated current imaging studies[69,70]. A 10-GHz bandwidth stand-alone graphene photo-detector for optical communication has recently been reported[71]. However, its maximum responsivity is low (6.1 mA W^{-1}) because the detector is designed to absorb only a small percentage of normal incident light. The responsivity can be improved by integrating the waveguide with the photo-detector because the light–graphene interaction length increases as light propagates along graphene's plane.

In a photo-detector with an integrated waveguide, optical signals from a waveguide can be transferred to a detector by either butt-coupled or evanescent-coupled schemes[72]. In the butt-coupled scheme, the detector and the waveguide are directly connected to each other on the same plane, whereas in the evanescent scheme the detector is on top of the waveguide. Our proposed approaches for both schemes are shown in Fig. 3 with their longitudinal and cross-sectional electric-field profiles. For the butt-coupling approach (Fig. 3a–c), graphene is sandwiched between half-etched silicon and polycrystalline silicon to minimize light reflection at the interface. In the evanescent-coupling approach (Fig. 3d–f), graphene is too thin to shift the light mode coming from the silicon waveguide and hardly absorbs any light. The polycrystalline silicon layer in Fig. 3b enables graphene to absorb light by changing the refractive index such that the mode shifts upwards.

To determine whether these detectors could adequately absorb light despite their extreme thinness, normalized detector lengths for the same absorption have been compared based on detector volume and electric-field intensity profiles. To achieve the same absorption as that of the widely used evanescent-coupled germanium detector, the evanescent-coupled graphene detector needs to be 3.5 times longer. In the butt-coupled case, however, the graphene detector only needs to be 12% longer because the graphene overlaps with the highest electric field. A further decrease in detector length can be achieved by using multilayer graphene because the absorption increases linearly with thickness for up to six layers[73].

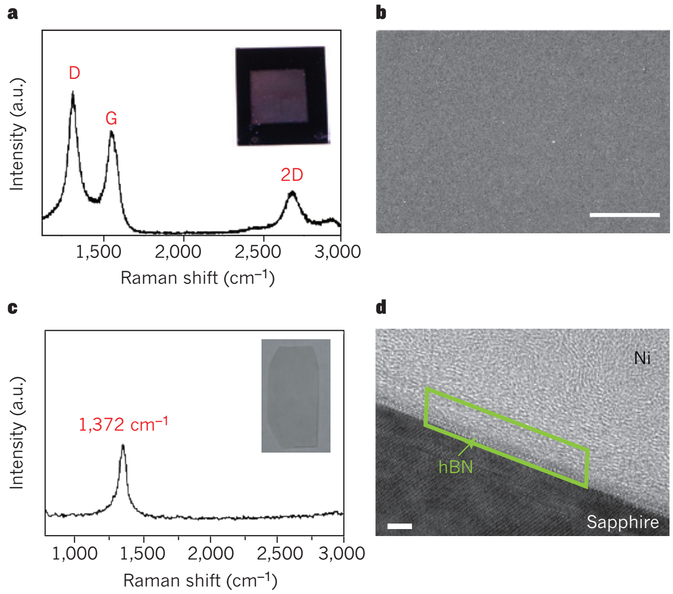


Figure 4 | Direct growth of graphene and hBN. **a**, **b**, The growth of graphene on hBN film showing the Raman spectrum (**a**) and a scanning electron microscopy image (**b**). Scale bar, 1 μm. The hBN film was grown on copper foil and transferred to a SiO_2/Si wafer. The inset in **a** shows a graphene sample grown on a 1 cm × 1 cm hBN film on a 2 cm × 2 cm SiO_2/Si wafer. Graphene shows typical Raman peaks of D (1,346 cm^{-1}), G (1,582 cm^{-1}) and 2D (2,700 cm^{-1}) and a homogeneous film structure with no particulate impurities. The G peak originates from the doubly degenerated vibrational mode of hexagonally structured graphene. The D peak is a disorder-induced band observed in a defective graphene structure. The 2D peak is the second order of the D band. **c**, **d**, The growth of hBN film on Al_2O_3 (0001) substrate showing the Raman spectrum (**c**) and a transmission electron microscopy image (**d**). Scale bar, 2 nm. The inset in **c** shows an hBN sample grown on the Al_2O_3 (0001) substrate measuring 1 cm × 2 cm. The hBN film shows a typical Raman peak at 1,372 cm^{-1}. a.u., arbitrary units.

To increase responsivity, effective extraction of photo-generated carriers should also be considered. Because doping control is difficult, graphene requires different carrier-extraction structures from semiconductor photo-detectors. A novel system for using metal-induced, built-in potential has been studied[35,69,70,74] and the device demonstrated[8,71]. Assuming light absorption of 2.3%, about 21% carrier extraction is estimated from 6.1 mA W^{-1} responsivity[71], and up to 60% carrier extraction can be anticipated[69]. Further structural optimization and graphene doping studies[75,76] are expected to increase the extraction efficiency.

Integration in semiconductor processes

Incorporating graphene into silicon devices is challenging, and careful selection of the right growth methods for the application is important. For graphene-based hybrid silicon-CMOS applications, graphene films should be deposited on topological surfaces as determined by the device architecture. There are generally two methods of graphene deposition: epitaxial growth on SiC(0001) by the sublimation of silicon atoms[77,78], and growth on catalytic metal films with subsequent transfer onto target substrates (CVD transfer)[79–81].

CVD transfer processes have already been demonstrated at wafer scales[82]. This method allows graphene to be transferred onto any type of material as long as it is flat, and there are no limitations on process temperatures. However, it is difficult to avoid defects such as holes, cracks and folds during the transfer[79–81]. Such defects would be minimized if graphene could be grown directly on underlying materials with topological surfaces at reasonable process temperatures.

The structures and process requirements of devices discussed in this Review are summarized in Fig. 5. For the radio-frequency transistor shown in Fig. 5a, for example, graphene will be deposited on hBN. However, growing hBN presents a challenge because growth could take place on two or more different materials such as metals (usually copper) and low-κ dielectrics. The problem is the high processing temperature for graphene and hBN, which exceeds the back-end-of-line processing temperatures of less than 500 °C. The most sensible way of preparing graphene and hBN is therefore by CVD on metal catalysts and film transfer.

An optical modulator with a buried architecture (Fig. 5b) requires multiple stacking of graphene and hBN (hBN/graphene/hBN/graphene/hBN) on Si and SiO_2. Because multiple transfers could increase the number of transfer-related defects, the entire film stack could be divided into two different stacks, which would require only a two-step transfer. The first stack would consist of graphene/hBN, and the second would be hBN/graphene/hBN. In this case, these stacks could be grown on catalytic metal film by CVD.

For the ridge structures shown in Fig. 5b, c, conformal stacking of graphene and hBN is needed on a topological silicon surface: hBN/graphene/hBN/graphene/hBN for optical modulators (Fig. 5b) and hBN/graphene/hBN for the photo-detector (Fig. 5c). Because the transfer method is limited to flat surfaces, transfer-free deposition on non-catalytic surfaces such as in silicon, graphene and hBN is required. Moreover, graphene and hBN depositions on non-catalytic surfaces must be prepared below 1,000 °C to avoid structural changes in CMOS devices, such as in doping profiles. There are two approaches for the transfer-free deposition of graphene on non-catalytic surfaces: on a catalytic metal film/substrate followed by eliminating the underlying metal[83–85], and directly on non-catalytic substrates by CVD[86,87]. Both approaches pose challenges. In the first approach, graphene can be torn during the elimination of the underlying metal. The second approach is not applicable to CMOS structures because of the high growth temperatures required (above 1,400 °C). Moreover, this latter process produces graphene flakes and dots, rather than continuous films. Therefore, direct growth of defect-free graphene at low temperatures on non-catalytic surfaces is required for CMOS integration. We have demonstrated that graphene film can be formed on a large-area hBN film by flow control of the precursor gas. Figure 4a shows the experimental results demonstrating that few-layered graphene can be directly grown on hBN at 1,000 °C using the CVD method. Graphene growth on the non-catalytic hBN film is enabled by enhancing the interactions of the precursor gas (C_2H_2) with the hBN film and by optimizing the flow direction as well as the pressure. This indicates that CVD has the potential to grow graphene on hBN directly; however, defects need to be minimized, and control of the number of layers should be improved.

For hBN growth, most studies have focused on catalytic metal substrates[88–90]. To realize the suggested modulator and detector structures, however, hBN on non-catalytic surfaces is required. We have attempted to deposit hBN directly on non-catalytic surfaces such as Si (001) and Al_2O_3 (0001) by thermal CVD. The hBN on Si could not be grown at all. However, a crystalline hBN has been successfully grown on Al_2O_3. Figure 4d shows a four-layered hBN on Al_2O_3 at 1,000 °C, achieved by controlling the flow kinetics where a large amount of precursor gas flows vertically relative to the substrate. These preliminary results suggest that the kinetic interactions of precursors with the substrate, as well as the chemical reactivity of surface atoms on the substrate, are important control parameters. Therefore, hBN could be deposited on Si if the Si surface is modified by adsorption of Al atoms, enhancing precursor reactivity with the surface.

To integrate graphene in silicon CMOS devices, direct growth of graphene on three-dimensional structures, simultaneous deposition on various underlying materials, and low processing temperatures are required. It will also be important to pursue further work on hBN film growth that could be used as dielectrics in radio-frequency transistors and as insulating materials in optical modulators and detectors. In-depth study of nucleation and growth mechanisms on non-catalytic surfaces will also accelerate the full realization of graphene-based hybrid silicon-CMOS applications.

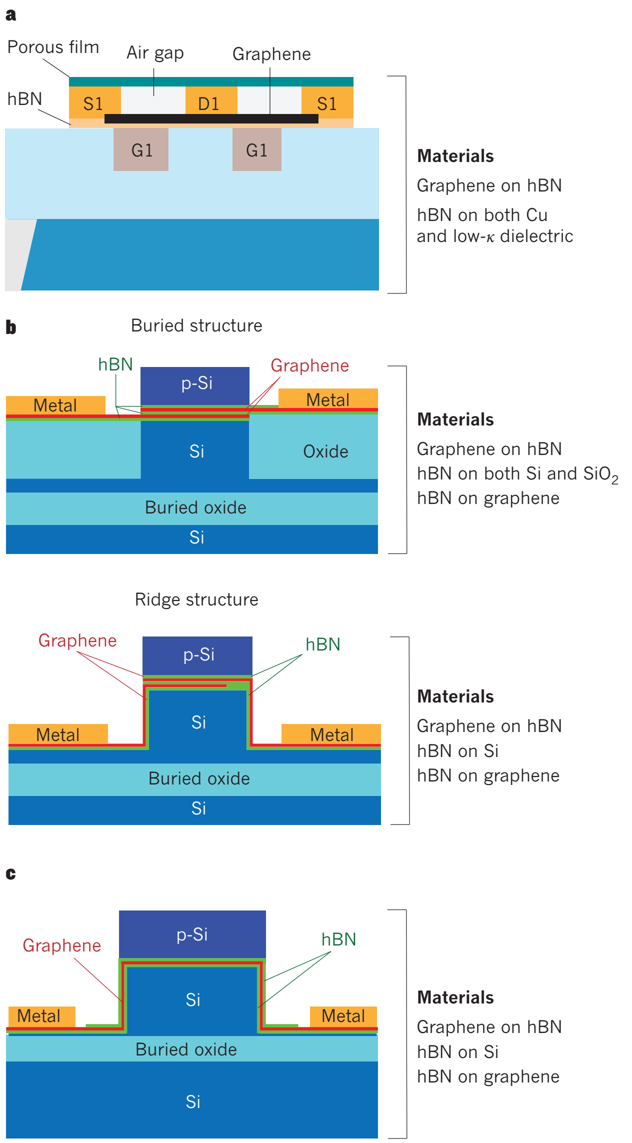


Figure 5 | Proposed device structures and process requirements for graphene and hBN films. **a**, Radio-frequency transistors have flat surface coverage. Film growth temperatures should not exceed the back-end-of-line processing temperatures around 500 °C. **b**, Optical modulators with buried structures also have flat surface coverage, whereas those with ridge structures have topological surface coverage. Both types have structural stability even at high fabrication temperatures of up to 1,000 °C. **c**, Photo-detectors have a ridge structure, giving them topological surface coverage. Like optical modulators, they have structural stability and are able to withstand fabrication temperatures of up to 1,000 °C.

Outlook

We have reviewed graphene devices for possible applications of current CMOS technology, focusing on radio-frequency transistors, optical devices and the deposition processes. By taking advantage of graphene's high carrier mobility and ultra-wideband optical absorption, we have proposed new architectures that are most likely to provide the earliest applications: the 'nothing-on-graphene' architecture for maximizing f_T, the graphene-gating optical modulator and the waveguide-integrated graphene photo-detector. We have also considered the limitations of graphene, such as the lack of bandgap through which f_{max} can be improved and its high contact resistance.

Graphene-based memory chips and microprocessors are unlikely to appear in the next few decades. In the meantime, graphene will have an important role in enhancing the performance of, and adding analogue functions to, silicon-based CMOS devices. However, for use in commercial graphene device applications, fabrication must be simple, reproducible and compatible with existing semiconductor processes. This implies that many of the current practices of graphene production will have to be changed drastically. Direct growth of high-quality graphene onto commonly used materials in semiconductor processes, such as Si and SiO_2, would be desirable. Advances confined to graphene alone are probably insufficient; new underlying materials, such as hBN, and processes to obtain such materials must be developed to fully maximize graphene's potential.

As interest in graphene continues to increase, we are optimistic that these technical issues will be resolved. The first application for graphene will probably be the radio-frequency switch, which has already shown a potential for improved performance, followed by optical modulators and photo-detectors. The next breakthrough, when the bandgap energy can be increased and controlled, is likely to expand the versatility of hybrid silicon-CMOS applications, making it a vital part of silicon CMOS evolution. ■

1. Kim, K. From the future Si technology perspective: challenges and opportunities. *Tech. Digest Int. Electron Devices Meet.* **10,** 1–9 (IEEE, 2010).
2. Novoselov, K. S. *et al.* Two-dimensional atomic crystals. *Proc. Natl Acad. Sci. USA* **102,** 10451–10453 (IEEE, 2005).
3. Novoselov, K. S. *et al.* Two-dimensional gas of massless Dirac fermions in graphene. *Nature* **438,** 197–200 (2005).
4. Zhang, Y. *et al.* Experimental observation of the quantum Hall effect and Berry's phase in graphene. *Nature* **438,** 201–204 (2005).
5. Novoselov, K. S. *et al.* Electric field effect in atomically thin carbon films. *Science* **306,** 666–669 (2004).
This paper reports the discovery of graphene by deposition of a graphene monolayer on a silicon substrate and the measurement of its field effect mobility to demonstrate its potential for electronic application.
6. Bolotin, K. I. *et al.* Ultrahigh electron mobility in suspended graphene. *Solid State Commun.* **146,** 351–355 (2008).
7. Liu, M. *et al.* A graphene-based broadband optical modulator. *Nature* **474,** 64–67 (2011).
This paper demonstrates an optical modulator using graphene for the first time.
8. Xia, F. *et al.* Ultrafast graphene photodetector. *Nature Nanotechnol.* **4,** 839–843 (2009).
9. Geim, A. K. & Novoselov, K. S. The rise of graphene. *Nature Mater.* **6,** 183–191 (2007).
10. Liao, L. *et al.* High-speed graphene transistors with a self-aligned nanowire gate. *Nature* **467,** 305–308 (2010).
11. Avouris, P. *et al.* Graphene-based fast electronics and optoelectronics. *Tech. Digest Int. Electron Devices Meet.* **10,** 552–555 (IEEE, 2010).
12. Lee, J. *et al.* RF performance of pre-patterned locally-embedded-back-gate graphene device. *Tech. Digest Int. Electron Devices Meet.* **10,** 568–571 (IEEE, 2010).
13. Lee, S. *et al.* Record RF performance of 45-nm SOI CMOS technology. *Tech. Digest Int. Electron Devices Meet.* 568–571 (IEEE, 2007).
14. Lai, R. *et al.* Sub 50-nm InP HEMT device with F_{max} greater than 1 THz. *Tech. Digest Int. Electron Devices Meet.* 609–611 (IEEE, 2007).
15. Kim, D.-H. *et al.* 30-nm InAs pseudomorphic HEMTs on an InP substrate with a current-gain cutoff frequency of 628 GHz. *IEEE Electron. Device Lett.* **29,** 830–833 (2008).
16. Singisetti, U. *et al.* Ultralow resistance *in situ* ohmic contacts to InGaAs/InP. *Appl. Phys. Lett.* **93,** 183502 (2008).
17. Liao, L. *et al.* Sub-100 nm channel length graphene transistors. *Nano Lett.* **10,** 3952– 3956 (2010).
This paper reports on a graphene transistor's performance, its feasibility and limits for radio-frequency applications, and suggests that graphene can outperform silicon-based radio-frequency transistors.
18. Schwierz, F. Industry-compatible graphene transistors. *Nature* **472,** 41–42 (2011).
19. Meric, I. *et al.* Channel length scaling in graphene field-effect transistors studied with pulsed current−voltage measurements. *Nano Lett.* **11,** 1093–1097 (2011).
20. Meric, I. *et al.* Graphene field-effect transistors based on boron nitride gate dielectrics. *Tech. Digest Int. Electron Devices Meet.* **10,** 556–559 (IEEE, 2010).
21. Lin, Y.-M. *et al.* 100-GHz Transistors from wafer-scale epitaxial graphene. *Science* **327,** 662 (2010).
22. Adam, S. *et al.* A self-consistent theory for graphene transport. *Proc. Natl Acad. Sci. USA* **104,** 18392–18397 (2007).
23. Robinson, J. *et al.* Contacting graphene. *Appl. Phys. Lett.* **98,** 053103 (2011).
24. Nagashio, N. *et al.* Metal/graphene contact as a performance killer of ultra-high mobility graphene analysis of intrinsic mobility and contact resistance. *Tech. Digest Int. Electron Devices Meet.* **9,** 565–568 (IEEE, 2009).
25. Bolotin, K. I. *et al.* Temperature-dependent transport in suspended graphene. *Phys. Rev. Lett.* **101,** 096802 (2008).
26. Hwang, E. H. *et al.* Carrier transport in two-dimensional graphene layers. *Phys. Rev. Lett.* **98,** 186806 (2007).

27. Nomura, K. *et al.* Quantum Hall ferromagnetism in graphene. *Phys. Rev. Lett.* **96,** 256602 (2006).
28. Chen, J.-H. *et al.* Intrinsic and extrinsic performance limits of graphene devices on SiO_2. *Nature Nanotechnol.* **3,** 206–209 (2008).
29. Isigami, M. *et al.* Atomic structure of graphene on SiO_2. *Nano Lett.* **7,** 1643–1648 (2007).
30. Meyer, J. C. *et al.* The structure of suspended graphene sheets. *Nature* **446,** 60–63 (2007).
31. Deshpande, A. *et al.* Spatially resolved spectroscopy of monolayer graphene on SiO_2. *Phys. Rev. B* **79,** 205411 (2009).
32. Dean, C. R. *et al*. Boron nitride substrates for high-quality graphene electronics. *Nature Nanotechnol.* **5,** 722–726 (2010).
33. Xue, J. *et al.* Scanning tunnelling microscopy and spectroscopy of ultra-flat graphene on hexagonal boron nitride. *Nature Mater.* **10,** 282–285 (2011).
34. Jeon, I. *et al.* Passivation of metal surface states: microscopic origin for uniform monolayer graphene by low temperature chemical vapor deposition. *ACS Nano* **5,** 1915–1920 (2011).
35. Giovannetti, G. *et al.* Doping graphene with metal contacts. *Phys. Rev. Lett.* **101,** 026803 (2008).
36. Xia, F. *et al.* The origins and limits of metal–graphene junction resistance. *Nature Nanotechnol.* **6,** 179–184 (2011).
37. Heersche, H. B. *et al*. Bipolar supercurrent in graphene. *Nature* **446,** 56–59 (2007).
38. Russo, S. *et al.* Contact resistance in graphene-based devices. *Physica E* **42,** 677–679 (2010).
39. Grosse, K. L. *et al*. Nanoscale Joule heating, Peltier cooling and current crowding at graphene–metal contacts. *Nature Nanotechnol.* **6,** 287–290 (2008).
40. Han, M. Y., Ozyilmaz, B., Zhang, Y. & Kim, P. Energy band-gap engineering of graphene nanoribbons. *Phys. Rev. Lett.* **98,** 206805 (2007).
41. Chen, Z., Lin, Y. M., Rooks, M. J. & Avouris, P. Graphene nano-ribbon electronics. *Physica E* **40,** 228–232 (2007).
42. Li, X. *et al*. Chemically derived, ultrasmooth graphene nanoribbon semiconductors. *Science* **319,** 1229–1232 (2008).
43. Xia, F. *et al.* Graphene field-effect transistors with high on/off current ratio and large transport band gap at room temperature. *Nano Lett.* **10,** 715–718 (2010).
44. Kim, M. *et al.* Fabrication and characterization of large-area, semiconducting nanoperforated graphene materials. *Nano Lett.* **10,** 1125–1131 (2010).
45. Liang, X. *et al.* Formation of bandgap and subbands in graphene nanomeshes with sub-10 nm ribbon width fabricated via nanoimprint lithography. *Nano Lett.* **10,** 2454–2460 (2010).
46. Bai, J. *et al.* Graphene nanomesh. *Nature Nanotechnol.* **5,** 190–194 (2010).
47. Palacios, T. *et al.* Applications of graphene devices in RF communications. *IEEE Commun. Mag.* **48,** 122–128 (2010).
48. Wang, H. *et al.* Graphene frequency multiplier. *IEEE Electron Device Lett.* **30,** 547–549 (2009).
49. Lin, Y.-M. *et al.* Wafer-scale graphene integrated circuit. *Science* **332,** 1294–1297 (2011).
50. Benner, A. Optical interconnects for HPC. *Optoelectronics Indust. Dev. Assoc. Workshop* (OIDA, 2011).
51. Miller, D. A. B. Device requirements for optical interconnects to silicon chips. *Proc. IEEE* **97,** 1166–1185 (2009).
52. Shacham, A. *et al.* Photonic networks-on-chip for future generations of chip multi-processors. *IEEE Trans. Comput.* **57,** 1246–1260 (2008).
53. Nairl, R. R. *et al.* Fine structure constant defines visual transparency of graphene. *Science* **320,** 1308 (2008).
54. Bonaccorso, F. *et al*. Graphene photonics and optoelectronics. *Nature Photonics* **4,** 611–622 (2010).
55. Eigler, S. A new parameter based on graphene for characterizing transparent, conductive materials. *Carbon* **47,** 2936–2939 (2009).
56. Wang, F. *et al*. Gate variable optical transitions in graphene. *Science* **320,** 206–209 (2008).
57. Reed, G. T. *et al.* Silicon optical modulators. *Nature Photonics* **4,** 518–526 (2010).
58. Green, W. M. *et al*. Ultra-compact, low RF power, 10 Gb/s silicon Mach–Zehnder modulator. *Opt. Express* **15,** 17106–17113 (2007).
59. Liao, L. *et al.* 40 Gbit/s silicon optical modulator for high speed applications. *Electron. Lett.* **43,** 1196–1197 (2007).
60. Watts, M. R. *et al*. Silicon microdisk modulators and switches. *Proc. 5th IEEE Int. Conf. Group IV Photonics* 4–6 (IEEE, 2008).
61. Xu, Q. *et al.* 12.5 Gbit/s carrier-injection based silicon micro-ring silicon modulators. *Opt. Express* **15,** 430–436 (2007).
62. Feng, N. *et al.* 30 GHz Ge electro-absorption modulator integrated with 3 μm silicon-on-insulator waveguide. *Opt. Express* **19,** 7062–7067 (2011).
63. Tang, Y. 50 Gb/s hybrid silicon traveling wave electroabsorption modulator. *Opt. Express* **19,** 5811–5816 (2011).
64. Bae, S. *et al.* Roll-to-roll production of 30-inch graphene films for transparent electrodes. *Nature Nanotechnol.* **5,** 574–578 (2010).
65. Meric, I. *et al*. Current saturation in zero-bandgap, top-gated graphene field-effect transistors. *Nature Nanotechnol.* **3,** 654–659 (2008).
66. Yin, T. *et al.* 31 GHz Ge n-i-p waveguide photodetectors on silicon-on-insulator substrate. *Opt. Express* **15,** 13965–13971 (2007).
67. Vivien, L. *et al.* 42 GHz p.i.n germanium photodetector integrated in a silicon-on-insulator waveguide. *Opt. Express* **17,** 6252–6257 (2009).
68. Ishikawa, Y. Near-infrared Ge photodiodes for Si photonics: operation frequency and an approach for the future. *IEEE Photonics J.* **2,** 306–320 (2010).
69. Park, J. *et al.* Imaging of photocurrent generation and collection in single-layer graphene. *Nano Lett.* **9,** 1742–1746 (2009).
70. Xia, F. *et al.* Photocurrent imaging and efficient photon detection in a graphene transistor. *Nano Lett.* **9,** 1039–1044 (2009).
71. Mueller, T. *et al.* Graphene photodetectors for high-speed optical communications. *Nature Photonics* **4,** 297–301 (2010).
This paper demonstrates a graphene photo-detector in a 10 Gbit s^{-1} optical data link for the first time.
72. Assefa, S. L. *et al.* CMOS-integrated optical receivers for on-chip interconnects. *IEEE J. Sel. Top. Quant. Electron*. **16,** 1376–1385 (2010).
73. Casiraghi, C. *et al.* Rayleigh imaging of graphene and graphene layers. *Nano Lett.* **7,** 2711–2717 (2007).
74. Mueller, T. *et al.* Role of contacts in graphene transistors: a scanning photocurrent study. *Phys. Rev. B* **79,** 245430 (2009).
75. Farmer. D. B. *et al.* Behavior of a chemically doped graphene junction. *Appl. Phys. Lett.* **94,** 213106 (2009).
76. Brenner. K. *et al.* Single step, complementary doping of graphene. *Appl. Phys. Lett.* **96,** 063104 (2010).
77. Berger, C. *et al*. Electronic confinement and coherence in patterned epitaxial graphene. *Science* **312,** 1191–1196 (2006).
78. Emtsev, K. V. *et al*. Towards wafer-size graphene layers by atmospheric pressure graphitization of silicon carbide. *Nature Mater.* **8,** 203–207 (2009).
79. Reina, A. *et al.* Large area, few-layer graphene films on arbitrary substrates by chemical vapor deposition. *Nano Lett.* **9,** 30–35 (2009).
80. Kim, K. S. *et al.* Large-scale pattern growth of graphene films for stretchable transparent electrodes. *Nature* **457,** 706–710 (2009).
81. Li, X. *et al.* Large-area synthesis of high-quality and uniform graphene films on copper foils. *Science* **324,** 1312–1314 (2009).
This report showed the possibility of growing a high-quality single layer of graphene (more than 95%) by CVD on copper foil.
82. Lee, Y. *et al.* Wafer-scale synthesis and transfer of graphene films. *Nano Lett.* **10,** 490–493 (2010).
83. Levendorf, M. P. *et al*. Transfer-free batch fabrication of single layer graphene transistors. *Nano Lett.* **9,** 4479–4483 (2009).
84. Hofrichter, J. *et al.* Synthesis of graphene on silicon dioxide by a solid carbon source. *Nano Lett.* **10,** 36–42 (2010).
85. Ismach, A. *et al.* Direct chemical vapor deposition of graphene on dielectric surfaces. *Nano Lett.* **10,** 1542–1548 (2010).
86. Park, J. *et al.* Epitaxial graphene growth by carbon molecular beam epitaxy (MBE). *Adv. Mater.* **22,** 4140–4145 (2010).
87. Zhang, L. *et al.* Catalyst-free growth of nanographene films on various substrates. *Nano Res.* **4,** 315–321 (2011).
88. Song, L. *et al.* Large scale growth and characterization of atomic hexagonal boron nitride layers. *Nano Lett.* **10,** 3209–3215 (2010).
89. Shi, Y. M. *et al.* Synthesis of few-layer hexagonal boron nitride thin film by chemical vapor deposition. *Nano Lett.* **10,** 4134–4139 (2010).
90. Liu, Z. *et al.* Direct growth of graphene/hexagonal boron nitride stacked layers. *Nano Lett.* **11,** 2032–2037 (2011).

Acknowledgements We thank U.-I. Chung, H. Kim, Y. Y. Lee, S. Jeon and S.-H. Lee for assisting with scientific discussions and contributions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to K.K. (kn_kim@samsung.com).

doi:10.1038/nature10681

Embracing the quantum limit in silicon computing

John J. L. Morton[1,2], Dane R. McCamey[3], Mark A. Eriksson[4] & Stephen A. Lyon[5]

Quantum computers hold the promise of massive performance enhancements across a range of applications, from cryptography and databases to revolutionary scientific simulation tools. Such computers would make use of the same quantum mechanical phenomena that pose limitations on the continued shrinking of conventional information processing devices. Many of the key requirements for quantum computing differ markedly from those of conventional computers. However, silicon, which plays a central part in conventional information processing, has many properties that make it a superb platform around which to build a quantum computer.

Technological progress has often faced seemingly fundamental barriers. When viewed from a new perspective, these barriers can be transformed into opportunities for innovation. So it is with the quantum mechanical limitations that silicon-based electronics have been charging towards over the past few decades. The silicon-processing length scale has shrunk tenfold every 15 years since 1971 and now stands at 22 nm. At such dimensions, there is a discrete, countable number of donors (10–100) in the transistor channel, and the electrical characteristics of individual dopant atoms can be observed in commercial field-effect transistors, albeit at low temperatures[1]. Furthermore, quantum effects such as the onset of quantum tunnelling of electrons through potential barriers limit the ability to confine charges to a densely packed array.

One way to address the quantum limit is to turn these properties into advantages: that is, to build a device whose function relies precisely on coherent quantum behaviour, including effects such as tunnelling, and to use the states of individual atoms, either natural or artificial, to store information. Such a device would be able to process quantum information[2] (a richer form of information than the ones and zeroes of classical bits) and would thus be capable of fundamentally outperforming conventional computers at solving certain classes of problem[3] (Fig. 1). In conventional electronics, charge is used to represent information; however, it is difficult to maintain coherent superpositions of charge states for long periods of time[4]; so, using charge to store information is a challenging approach for quantum computation. By contrast, the spin of an electron can exist in coherent states for periods of as long as seconds[5]. This is many orders of magnitude longer than the few nanoseconds that it takes to manipulate a spin, in principle allowing quantum information to be perpetually maintained through quantum error-correction codes[6].

Silicon is a particularly attractive material for hosting spin quantum bits (qubits) for several reasons. Silicon has low spin–orbit coupling, and silicon's nuclear-spin-bearing isotope has a low natural abundance (only 5% of natural silicon is ^{29}Si, which can be removed by isotopic enrichment[7]). Both of these properties contribute to long spin coherence times. Furthermore, the advanced state of silicon electronics offers a common platform for spin qubits alongside sophisticated classical integrated circuits.

Spin qubits can be realized in silicon using naturally confined donor-bound spins or lithographically defined silicon-based quantum dots (Box 1 and Fig. 2). In this Review, we evaluate how such approaches perform with respect to key metrics, which are as applicable to quantum computers as they are to conventional ones: robustness against information loss, speed and fidelity of control operations and measurement, and interconversion between different storage media. Finally, we assess the challenges that lie ahead for silicon-based spin qubits in terms of materials, as well as the nanofabrication requirements that must be met to construct a quantum computer in silicon.

Spin coherence and mechanisms for decoherence

The corruption of the state of the electron spin is characterized by two timescales: T_1 describes spin relaxation from the spin-up to the spin-down state (or the loss of classical information stored in the spin), whereas T_2 is the coherence time, which characterizes spin decoherence (or the loss of the phase information that is necessary to store quantum information) (Fig. 3). Silicon was one of the first semiconductors to be highly purified and grown in large crystals, which led to it being used to explore and develop new electron spin-resonance (ESR) methods.

For electrons bound to donors in silicon, T_1 becomes very long (minutes to hours[8]) at low temperatures. Above a few kelvin, T_1 is limited by thermal excitation to the valley–orbit excited state of the donor, whereas it is controlled by phonon scattering at lower temperatures. For donors, T_2 can be considerably shorter than T_1, and measuring T_2 necessitated the development of electron spin-echo techniques. In the first microwave electron spin-echo experiment, in 1958, Gordon and Bowers measured the decay of the spin echo from electrons bound to lithium and phosphorus donors in silicon[9]: a T_2 of 0.5 ms was found for a crystal of isotopically enriched ^{28}Si:P. This T_2 was longer than that observed in natural Si, thus demonstrating that ^{29}Si contributes to the electron spin decoherence, and was probably the longest coherence time measured for an electron spin in a condensed matter system for almost half a century.

As interest grew in identifying electron spin systems with long coherence times for applications in quantum information processing, understanding the limits of electron spin coherence took on a new importance. Studies of bulk silicon doped with donors have established that interactions among the donor electron spins and between the electrons and residual ^{29}Si nuclear moments were responsible for the T_2

[1]Department of Materials, University of Oxford, Parks Road, Oxford OX1 3PH, UK. [2]Clarendon Laboratory, University of Oxford, Parks Road, Oxford OX1 3PU, UK. [3]School of Physics, University of Sydney, New South Wales 2006, Australia. [4]Department of Physics, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [5]Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544, USA.

BOX 1

Confining electron spins in silicon

Spin qubits can be realized in silicon in two ways: in donor-bound spins in silicon or in silicon-based quantum dots.

Donors

Semiconductors provide a natural mechanism for trapping single spins: using electrons bound to individual donor atoms at low temperature. Phosphorus donors have been the most-studied dopant atom in silicon[8], especially for quantum information applications[12,70]. Other donors, including arsenic (ref. 37), antimony (ref. 35), bismuth (refs 52, 75) and lithium (ref. 76), may have different advantages. To incorporate donor-based spins into nanoscale devices, it is necessary to position electrostatic gates with respect to the donor and to position many donors with respect to each other (the precision that is required depends on the precise architecture that is sought[25,70,77–79]).

Donor atoms can be positioned in silicon by using ion implantation, even at the level of individual donors[80,81], although the precision with which this can be achieved is limited to approximately 10 nm owing to straggling. Transport measurements — in which electrons tunnel through individual implanted donor atoms in metal–oxide–semiconductor field-effect transistor (MOSFET)-like devices[82] or in fin field-effect transistors (FinFETs)[83] — can be performed. Recently, it was shown that donors can be tunnel-coupled directly to silicon single-electron transistors, an architecture that allows measurement with a large signal-to-noise ratio[42]. Moreover, donors can be positioned with extremely high precision by using a lithographic technique based on the removal of a hydrogen resist by a scanning tunnelling microscopy tip[67,84]. Lines of donors for use as leads and gates can also be fabricated in this way, allowing the precise alignment of a cluster of a few donors[74] or a single donor.

Quantum dots

Alternatively, spins can be confined in 'artificial atoms', by using lithographically defined quantum dots[71,85]. In contrast to donors, quantum dots are extremely tunable[86,87], but they have only recently been developed in silicon because of materials challenges that ranged from the relatively large effective mass of electrons in silicon to the substantial mismatch in the lattice constants of various group IV semiconductors. These challenges have been overcome, and quantum dots have been formed in Si/SiGe heterostructures using Schottky gates[88,89], as well as in structures that closely resemble silicon MOS transistors[65,90,91], in patterned regions of ultra-high dopant density[74,92] and in gated nanowires[93,94]. In each case, individual charge occupation has been demonstrated[94–96].

An early concern about silicon-based quantum-dot spin qubits involved the conduction band in silicon, which, in its unstrained bulk form, has six equivalent minima (valleys). Early experiments on two-dimensionally confined systems showed that, although it was relatively easy to move four of the valley states to a high energy, the quantum states arising from the two *z*-valleys, perpendicular to the two-dimensional layer, were nearly degenerate unless very large magnetic fields were applied[97,98]. Valley degeneracy could be a useful property of qubits[99]; however, in current experimental systems, it is thought that this degeneracy of the *z*-valley states is lifted by coupling between them[100], as demonstrated by recent observations of spin blockade in silicon-based quantum dots[45,46,101].

value measured by Gordon and Bowers. Spectral diffusion of electron spins resulting from their interaction with nuclear magnetic moments has been observed in many systems, but recent theoretical developments using cluster expansion techniques have shown excellent agreement with electron spin-coherence measurements in silicon[10,11]. By reducing the ^{29}Si concentration to 50 p.p.m. and the donor density to $1\times10^{14}\,\mathrm{cm}^{-3}$, as well as by using techniques to suppress the effects of dipolar interactions among the donors[5,12], coherence times of about 10 s have been observed at 1.8 K. This T_2 is still two orders of magnitude lower than T_1 at this temperature, and efforts to further reduce the effects of nuclear moments and interactions among the donor electron spins are under way.

Donor-bound electrons near the surface of the silicon can have significantly shorter spin-coherence times than those in the bulk: Schenkel and colleagues[13] found a T_2 of 2.1 ms for donors with a distribution that peaked at a depth of 150 nm from the silicon surface, and this decreased to about 0.75 ms for donors that were 50 nm below the surface. In each case, T_1 was about 15 ms at 5.2 K, which was comparable to that of bulk antimony donors. The origin of this additional decoherence is under investigation: it is possible that it arises from interactions with spins at the silicon surface or from defects remaining as a result of the ion implantation of the dopant atoms. Charge noise associated with the surface could also couple to the donor spins by Stark shifting the electron spin-resonance frequency.

Although the coherence of electrons that are tightly bound to donors extends to seconds, the situation is different for free electrons in silicon. High-mobility two-dimensional electron systems in Si/SiGe heterostructures show both a T_1 and a T_2 in the microsecond range at 4.2 K, with a T_2 of about 3 μs being the longest directly measured coherence time[14]. These results are comparable to those obtained at higher temperatures from coherent spin transport experiments in bulk silicon[15]. A T_1 of about 1 ms for free electrons in a Si/SiGe two-dimensional electron gas has been inferred from high-field electrically detected magnetic resonance experiments[16], but this has not been confirmed by direct pulsed measurements.

In pulsed-ESR measurements of a two-dimensional electron system, it has been found that T_2 can exceed T_1 (in general, T_2 is limited to $2T_1$ and only reaches this length when the dominant process is relaxation). Values of T_1 and T_2 in the microsecond range can be understood as arising from the spin–orbit interaction in the form of a Rashba effective magnetic field[14,17]. Lower-mobility two-dimensional electrons in silicon metal–oxide–semiconductor (MOS) structures have a T_1 and T_2 that are about an order of magnitude less than in high-mobility Si/SiGe heterostructures[18].

For quantum dots, T_1 times of many milliseconds and even seconds have been reported for both MOS[19] and Si/SiGe-based[20,21] systems, although little is known about the coherence of these spins. At present, several research groups are making an intense effort to determine the dephasing time, T_2^*, in electron double-dot experiments in silicon. Such measurements build on techniques used to determine T_2^* for GaAs quantum dots[22]. There has been one pulsed-ESR measurement of isolated electrons localized at a MOS interface by disorder potentials (so-called natural quantum dots); this gave a lower bound for T_2 of about 30 μs at 350 mK (ref. 18), with this short T_2 being attributed to exchange coupling between electrons in the shallow quantum dots.

Long coherence times are advantageous for quantum information processing. However, some form of error correction, analogous to refreshing dynamic random access memory (DRAM) bits in classical computing, will always be essential. A key figure of merit for fault-tolerant quantum computation therefore becomes the ratio of the coherence time to the duration of a qubit control operation.

Control of electron spins in silicon

One of the advantages of encoding quantum information in electron spins is the ability to manipulate spin by using established ESR techniques[23], in which resonant microwave fields drive spins between eigenstates, typically on timescales of 1–100 ns (Fig. 4a). By applying

a Classical information processing

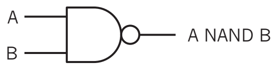
Bit

0 or 1

Boolean logic

NOT AND (NAND)

A	B	A NAND B
0	0	1
0	1	1
1	0	1
1	1	0



b Quantum information processing

Qubit

$\psi = \alpha|0\rangle + \beta|1\rangle$

Quantum logic

Controlled NOT (CNOT)

Before	After
$\lvert$Control Target$\rangle$	$\lvert$Control Target′$\rangle$
$\lvert 00\rangle$	$\lvert 00\rangle$
$\lvert 01\rangle$	$\lvert 01\rangle$
$\lvert 10\rangle$	$\lvert 11\rangle$
$\lvert 11\rangle$	$\lvert 10\rangle$

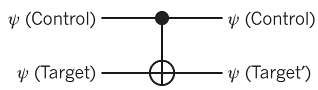


Figure 1 | Classical versus quantum information. The classical binary digit (bit) exists in one of two states, known as 0 and 1 (**a**). Quantum mechanics, by contrast, allows a two-state system to exist in a superposition of the two states, with a defined phase between the two (**b**). The most basic unit of quantum information (the quantum bit, or qubit) is therefore written as a combination of the states $|0\rangle$ and $|1\rangle$ defined by the complex numbers α and β. In both types of information processing (classical and quantum), there is a universal set of logic operations from which any algorithm can be composed. As is the case for conventional logic gates (such as NAND), the action of quantum logic operations (such as CNOT) can be understood from truth tables (centre), which show how output states depend on inputs. Importantly, quantum logic gates can act on superpositions of input states (bottom).

an appropriate sequence of pulses with a precise duration and phase, arbitrary single- or multiple-qubit operations can be performed. ESR techniques can also be used to implement dynamical decoupling schemes, which can be used to combat the effects of random environment variations, thus extending coherence times[24].

The spins in silicon provide well-defined, reproducible qubits, especially in the case of donors. However, as a result of this uniformity, globally applying ESR pulses to a large ensemble of qubits manipulates all of them in the same way. Some quantum computing architectures can function using only such 'global control' methods[25], but it is generally advantageous to be able to selectively address individual spins. There are two leading approaches to achieving this selectivity (Fig. 4b). In the first, the oscillating magnetic fields that are used to drive spin resonance are spatially restricted, which can be achieved, for example, by fabricating local resonators that are located close to each individual qubit. This is technically demanding for a.c. magnetic fields; however, a quantum dot can be driven using an a.c. electric field, by placing the dot in a magnetic field gradient (for example, provided by a nearby micromagnet), as was recently demonstrated on a III–V quantum dot[26].

To selectively manipulate tightly packed donor-based or quantum-dot-based qubits within a global a.c. magnetic field, it is possible to spatially distort the wavefunction of the electron, modifying its spin-resonance frequency and bringing it in or out of resonance with the applied field. Spatial distortion of the donor wavefunction can be achieved by applying electric fields (Stark shifting)[27] or by straining the lattice[28]; the latter case allows tuning of the hyperfine interaction strength by nearly 1%.

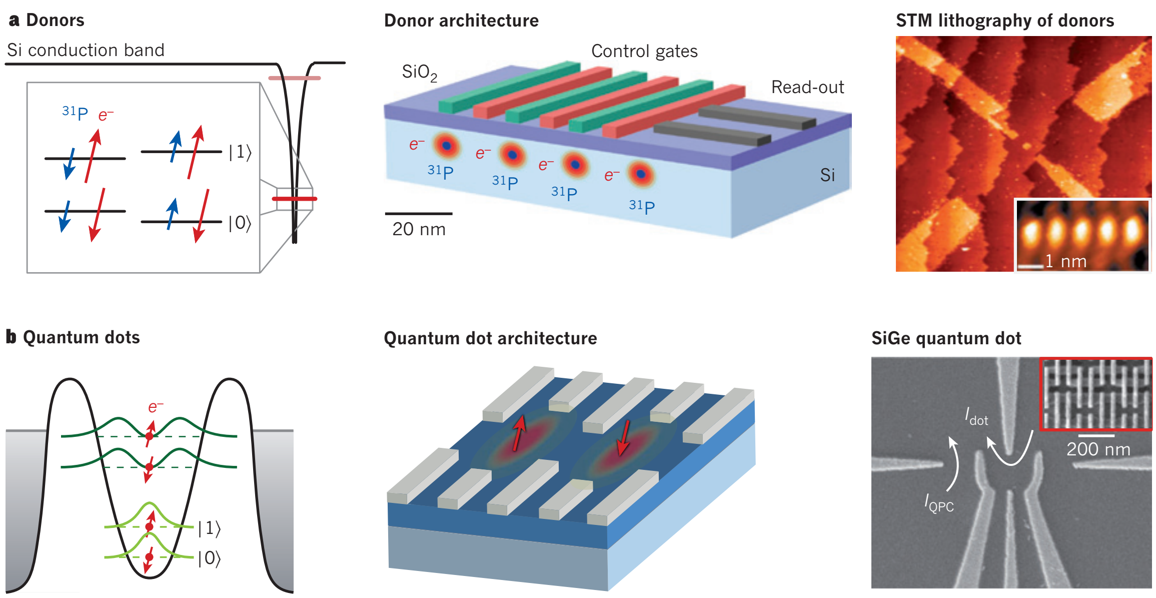


Figure 2 | Fundamental building blocks of silicon-based qubits. Electrons are localized using either donors (for example, phosphorus) (**a**) or artificial quantum dots (**b**) (see also Box 1). Left, The electron spin eigenstates (red), which are energetically separated by a Zeeman interaction with an external magnetic field, can be used to encode the $|0\rangle$ and $|1\rangle$ states of a quantum bit. Donor electron spins have an additional hyperfine interaction with the ^{31}P nuclear spin (blue), which can be exploited for both storing information in the nuclear spin and for read-out. Quantum-dot electrons have several bound orbital states (green). Centre, Quantum computing architectures usually incorporate buried arrays of donors or quantum dots with control gates on the surface of the silicon. Right, Scanning probe techniques, such as scanning tunnelling microscopy (STM), can be used to fabricate arrays of donors with atomic precision (top, inset), as well as the electrical leads and gates with which to control and address them (top, main image). Quantum dots are usually fabricated using SiGe or metal–oxide–semiconductor (MOS) structures, in which a buried two-dimensional electron gas is constricted to form a dot containing a single electron. The quantum dot can be measured using the current flowing through the dot (I_{dot}) and/or flowing through a nearby quantum point contact (I_{QPC}). A typical surface-gate-defined quantum dot (bottom) is shown on the same scale as the gate layers that are used in conventional 22-nm static random access memory (SRAM) chips (inset)[73]. The image on the right of panel **a** is reproduced, with permission, from ref. 74. The inset within this image is reproduced, with permission, from ref. 67.

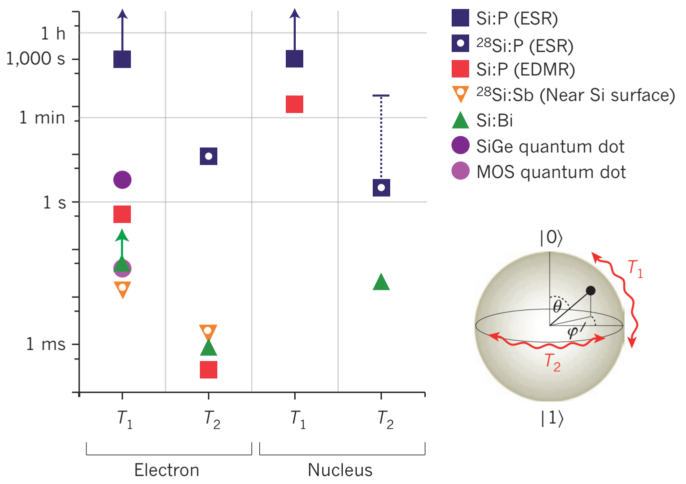


Figure 3 | Information lifetimes for various spins in silicon. A qubit can be represented as a point on the surface of a sphere (bottom right). Corruption of the qubit state can be separated into two timescales: T_1 describes the randomization of θ, whereas T_2 describes the randomization of φ. This figure summarizes the measurements that have been made for the T_1 and T_2 of the electron spin of donors in silicon and of silicon-based quantum dots. In the case of donor-bound electron spins, the relaxation times of the donor's nuclear spin can also be extracted. Unless otherwise indicated, natural silicon (with ~5% ^{29}Si) was used. The arrows indicate that these times become even longer at lower temperatures; the dashed line indicates unpublished results. From the top to bottom of the graph, electron T_1 data are taken from refs 8, 21, 41, 52 and 75, 19, and 13; electron T_2 data from refs 5, 13, 52 and 33; nuclear T_1 data from refs 8 and 57; and nuclear T_2 data from M. Thewalt (measured by optically detected magnetic resonance; personal communication) and ref. 51, and ref. 52. EDMR, electrically detected magnetic-resonance.

Read-out and measurement methods

In addition to the ability to control qubits, it is necessary to be able to determine their state after implementing an algorithm. An entire class of quantum information processing (using cluster states) can be implemented by global manipulation, if individual donors can be selectively read out. Superficially, read-out seems a simple problem: many thousands of spin-resonance experiments are undertaken every day using read-out by magnetic induction, and single spins can be measured in a similar way[29]. The difficulty arises in that the magnetic signal from a single spin is extremely small, requiring a long time for a single measurement. To overcome this challenge, alternative techniques have been developed. These are based on converting the information from the spin to another property that is more readily measurable, such as current (charge) or light.

Electrically and optically detected magnetic resonance

One class of measurement methods uses the spin of an electron to control its ability to move: the spin information is mapped onto a current through a process of spin-dependent scattering, tunnelling or recombination. Spin-dependent recombination was first demonstrated nearly 50 years ago[30], when the photoconductivity of silicon was modified by incoherently manipulating the spin of an ensemble of phosphorus donor electrons. More recently, this technique has been used to investigate ensembles of fewer than 100 donors[31] and extended to allow read-out after coherent spin manipulation[32], resulting in phase coherence times of ~100 μs (ref. 33). These times are shorter than those that can be measured by standard ESR methods because the free charge carriers (which form a fundamental part of these spin-dependent recombination methods) also induce spin relaxation. In addition, the recombination rate is generally fixed, resulting in either long read-out times or short state lifetimes, although it might be possible to control the recombination rate by varying the wavefunction overlap between the donor and probe spins using externally applied electric fields[34].

To overcome the limitations associated with photoexcited carriers, gated structures such as MOS field-effect transistors (MOSFETs) could be used to provide greater control over the conduction electrons used for read-out. The donors in such devices have been studied using spin-dependent scattering processes[35,36]. Indeed, the current that passes through a MOSFET can be controlled according to the state of the donor spins in the channel, providing a glimpse of the power of incorporating quantum systems into conventional electronics. At present, the spin-dependent effect on MOSFET conductivity is weak: changes of up to 0.03–0.3% were observed using high magnetic fields (3.4 T)[37]. Nevertheless, with sufficient signal averaging, it is possible to detect the signature of a single spin (of an unknown defect) in a MOSFET[38].

Another promising measurement avenue is the use of high-precision optical spectroscopy to selectively photoexcite donors in specific spin states, along the lines of optically detected magnetic resonance[39]. The photoexcitation of donors can then be detected by an optical pump-probe approach, by measuring the change in photoconductivity as above or by inducing scattering between quantum Hall edge states. Although the methods described in this section can be scaled to the single-spin level, a range of single-spin read-out schemes have no parallel for measuring ensembles and have only recently become experimentally accessible.

Single-shot read-out of electron spins

Single-shot read-out of single spins relies on a version of spin-to-charge conversion in which the spin controls the tunnelling of an individual electron, which can be measured with an integrated charge sensor[22,40] (Fig. 5). An electron is loaded in a quantum dot[21] or, equivalently, a donor[41] and occupies either a spin-up or spin-down state, and these states are Zeeman split by an applied magnetic field. A gate voltage is used to tune the quantum dot or donor so that the spin states straddle the Fermi level of a nearby charge reservoir. If the higher-energy spin state is occupied, the electron will tunnel off the donor or out of the quantum dot, and the resultant change in the charge state of the donor or quantum dot will be detected as a current transient. Shortly afterwards, an electron of opposite spin will tunnel into the lower-energy spin state, and the detectable current signature will disappear.

This method has recently been demonstrated in silicon for single electrons confined either to a single donor[42] or in a quantum dot[21]. Using this and similar methods, T_1 was measured for single spins in both silicon-based quantum dots and donors in silicon, and times up to and exceeding 1 s were recorded[19–21,42]. T_1 has been observed to increase extremely rapidly with decreasing strength of the magnetic field, which is in agreement with theory[43].

Spin blockade and singlet–triplet qubits

In low but non-zero magnetic fields, in which T_1 is longest, the Zeeman splitting that is used for single-shot read-out of electron spins is relatively small, leading to lower-fidelity measurements than in stronger magnetic fields. For this reason, it is interesting to consider encoding a qubit using two-electron spin-singlet and spin-triplet states[22,44]. The read-out mechanism in this approach relies on the energy splitting between the singlet and triplet states for two electrons in a single quantum dot (or on a single donor) to prevent tunnelling. This effect, known as Pauli spin blockade, has been observed in both silicon-based quantum dots[45,46] and donors in silicon[47].

Hybrid qubits in silicon

Current technologies transfer information between different physical representations to take advantage of their relative strengths: for example, moving information from magnetic domains of hard disks to charge in SRAM cells to light pulses in optical fibres. Similarly, one of the advantages of electron spin qubits is the flexibility they provide through their interactions with other quantum degrees of freedom, such

as nuclear spin, charge and photons, all of which can be used to improve coherence lifetimes, processing or read-out[48].

Because their magnetic moment is weaker than that of electron spins, nuclear spins typically have longer relaxation times (both T_1 and T_2)[49]; therefore, the nuclear spin of a donor provides a potentially long-lived memory element in which the electron spin state may be stored[50]. Through a combination of resonant microwave and radio frequency pulses, the state of a P donor's electron spin may be coherently swapped with that of the ^{31}P nuclear spin in 'read' and 'write' operations, with fidelities of more than 97% (ref. 51). By comparing the loss of coherence after a write-then-read operation, nuclear T_2 times of up to several seconds have been measured for ^{31}P. These times are comparable to the best times achieved for electron spins, but they should be more robust in proximity to interfaces and in the presence of residual ^{29}Si. Similar operations could be performed using other group V donors. Indeed, the large nuclear spin of ^{209}Bi

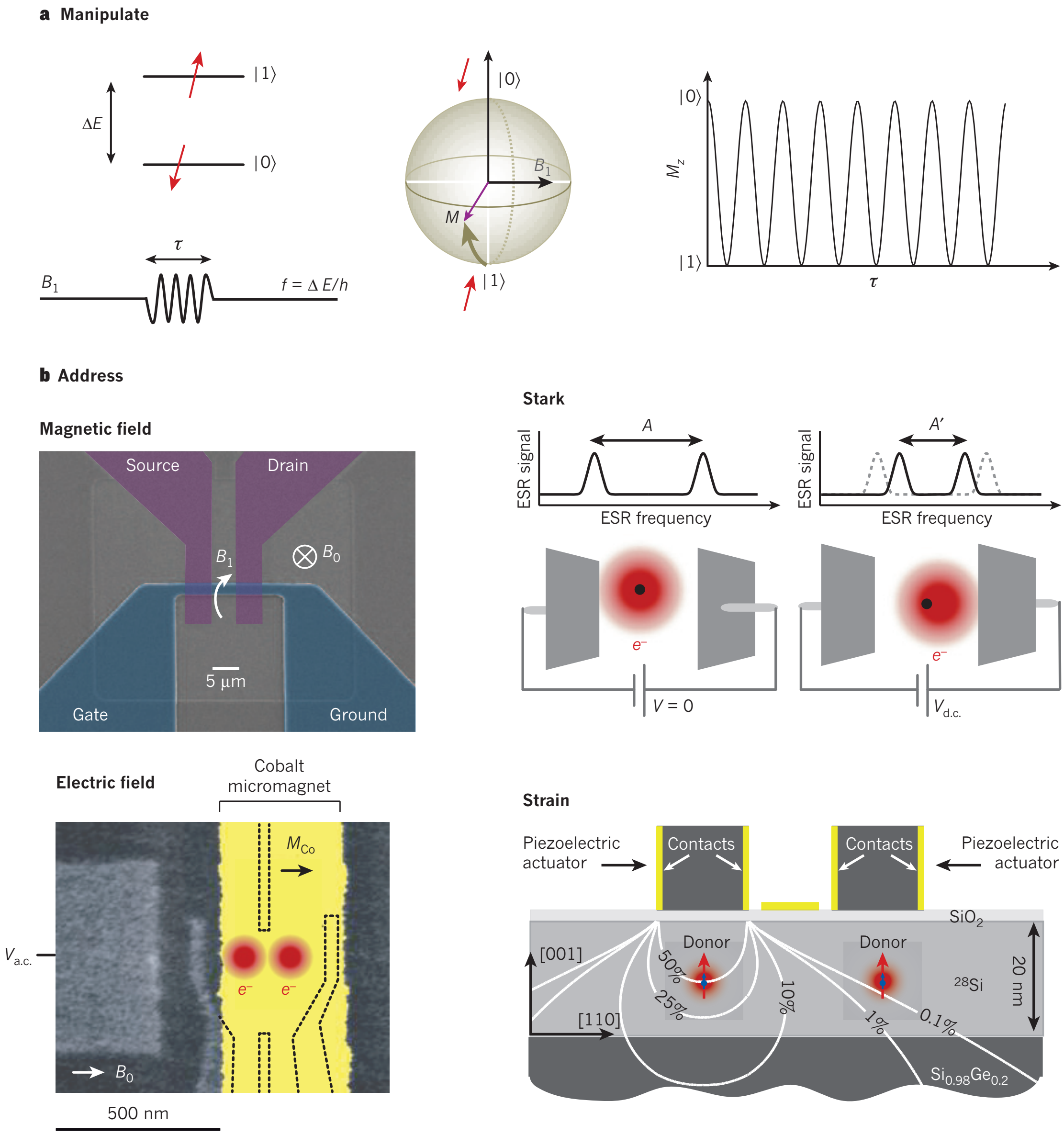


Figure 4 | Controlling and manipulating spins. **a**, Spins (whose eigenstates are separated in energy by ΔE) can be coherently manipulated using resonant microwave pulses of frequency, $f = \Delta E/h$, (where h is Planck's constant). A pulse of a given magnetic field strength, B_1, and duration, τ, will drive a spin from the $|0\rangle$ to the $|1\rangle$ state, whereas a pulse of half this duration will leave the spin in an equal superposition of the two states. The evolution of the spin magnetization, $\boldsymbol{M}$, under an applied pulse can be understood as a rotation around the Bloch sphere (green trajectory) perpendicular to the direction of B_1. This can be measured as coherent oscillations in the z component of the magnetization as the duration of the pulse is increased. **b**, These control pulses can be locally applied using microwave transmission lines located near each qubit, driving the spin directly with an a.c. magnetic field, B_1, perpendicular to the static field B_0 (top left). The blue structure functions as a gate that induces a two-dimensional electron gas, as well as a microwave transmission line. The current flowing through the source and the drain contacts is measured to perform electrically detected magnetic resonance of P donors under the gate. Alternatively, the spin of a quantum dot can be driven indirectly using an a.c. voltage, $V_{a.c.}$ (bottom left), which modulates the position of the spin within a magnetic field gradient created by a cobalt micromagnet (with magnetization M_{Co}). Spin manipulation pulses can also be applied globally, and the resonance of individual spins can be modified so that they can be selectively manipulated. This can be achieved by locally applying a d.c. voltage, $V_{d.c.}$, and exploiting the Stark effect (top right) or by using strain fields (bottom right). This modifies the wavefunction of the donor electron, changing the hyperfine interaction with the nuclear spin and thus the electron spin-resonance frequency. Strain is applied using piezoelectric actuators, which strain a ^{28}Si epitaxial layer on SiGe. The image at the top left of panel **b** is adapted, with permission, from ref. 36. The image at the bottom right of panel **b** is reproduced, with permission, from ref. 28. The image at the bottom left of panel **b** is reproduced, with permission, from ref. 26.

(I = 9/2, where I is the total angular momentum of the nucleus) provides a large Hilbert space in which several spin qubits could be stored[52]. Although the weak magnetic moment of the nuclear spin also makes it challenging to initialize qubits, nuclear spins can be polarized through coupling to an electron spin (for example, ^{31}P in Si), by using optically driven dynamic nuclear polarization[53,54] or spin-resonance pulse sequences[55].

In addition to advantages for quantum memory, nuclear spins provide a route to achieving quantum non-demolition measurement, which is an ideal quantum measurement that does not itself perturb the system beyond the inevitable collapse of the wavefunction that is brought about by discovering its state. In the scheme proposed by Sarovar and colleagues[56], a quantum non-demolition measurement of a nuclear spin is made by measuring the resonance frequency of a coupled electron that is shifted by the state of the nuclear spin as a result of the hyperfine coupling. The electron spin is measured using one of the methods described above, each of which is destructive to the state of the electron but largely preserves the state of a coupled nuclear spin[57] (Fig. 5c).

Of the various approaches for quantum computing using solid-state devices, researchers have made particularly strong progress in coupling multiple qubits together by using qubits based on superconducting circuits, generating up to three-qubit entangled states[58,59]. However, such qubits remain limited in their applicability by short coherence times, which are typically a few microseconds to tens of microseconds. One important application of spins in silicon could be as memory elements operating in the microwave regime: superconducting qubits can exchange their state with a microwave photon in a superconducting coplanar waveguide cavity, and the state of this photon can be stored, in turn, within a spin ensemble[60–63].

For long-distance transmission of quantum information, a similar state transfer would need to be performed with photons in the optical regime, allowing seamless quantum computation and communication. Unfortunately, the poor optical properties of silicon make this challenging. Optical emission from excitons bound to phosphorus donors in silicon can be used to measure the state of both the electron spin and the nuclear spin[54]; however, the measurement efficiency is low, prohibiting even single-spin measurements let alone qubit state transfer. Other defects in silicon, such as erbium, result in luminescence from an inner-shell transition and may be better suited to such applications. If such a state transfer is successful, single optical photons could be routed on-chip using silica-on-silicon waveguides, which have been used to perform quantum logic operations, albeit non-deterministically[64].

Materials and nanofabrication

Great progress has recently been made in the development of spin-based quantum devices in silicon, but there are a number of challenges to be met before these structures become practical, scalable systems. The most pressing issues revolve around materials: limiting the

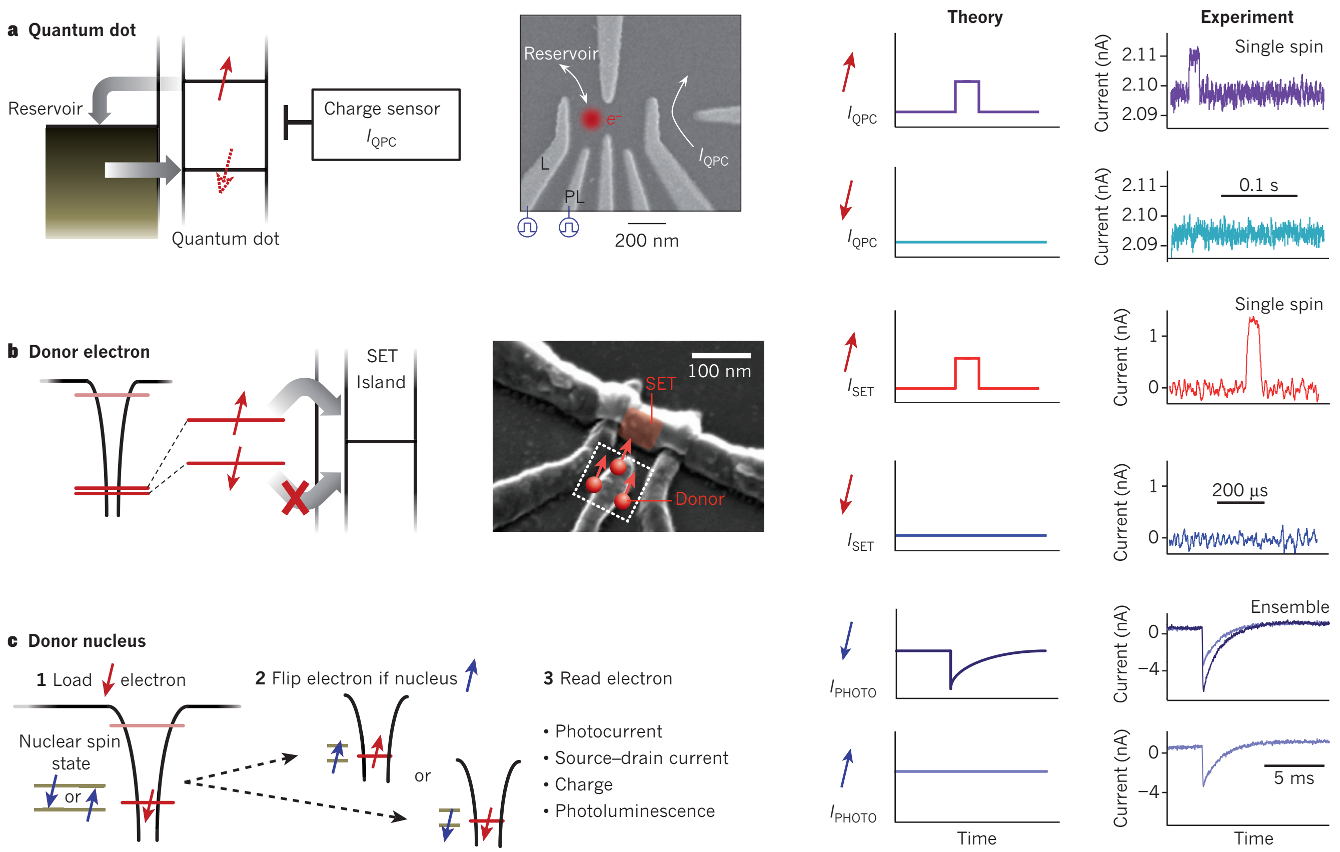


Figure 5 | Spin read-out of quantum dots and donors. Three read-out schemes, each using spin-controlled electron current are shown, including a schematic of how the spin-current conversion is performed (left), a scanning electron microscopy image of the device (centre) and data showing spin measurement (right). **a**, The current through a charge sensing quantum point contact (I_{QPC}), can be used to monitor the occupancy of a quantum dot. Selectively biasing the quantum dot (with a voltage pulse on the gates labelled L and PL) results in electron tunnelling of the dot to a nearby reservoir only if it is in the spin-up state. The signal from the charge sensor changes only if the electron moves. Experimental data are taken from ref. 21. **b**, Alternatively, an electron may be induced to tunnel from a donor onto the island of a nearby single electron transistor (SET), modifying its current. Experimental data are taken from ref. 42. The image in the centre of panel **b** is reproduced, with permission, from ref. 42. **c**, The state of the donor nuclear spin can also be measured as a result of its hyperfine coupling to the electron. For example, a charge-trap mechanism can be used; this results in a reduction in photocurrent only if the electron is spin down. By using the hyperfine interaction to selectively rotate the electron based on the spin of the nucleus, the current reflects the nuclear state. Experimental data are taken from ref. 57.

impact of impurities in the devices, controlling defect formation, and fabricating sufficiently small and reproducible devices. The easiest problem to solve is probably the requirement for isotopically enriched ^{28}Si. Processing ^{28}Si at the level needed for the Avogadro Project[7], or for multi-second T_2 times, is challenging, but a donor-electron spin coherence of 60 ms has been obtained with commercially grown epitaxial layers of ^{28}Si on natural Si at a moderate cost[12].

Patterning the gates that control the electrons and their spins at a sufficiently small length scale is often a significant hurdle in device research. The size of quantum-dot-based devices is much less than 1 μm, and donor-based devices are even smaller. It is tempting to look at what the semiconductor industry can produce (Fig. 2b, inset) and to declare victory — the industry already fabricates copious quantities of devices that are smaller than those in research laboratories. However, this ignores the fact that modern silicon devices are designed and optimized for operation at room temperature, whereas silicon-based quantum structures operate at millikelvin temperatures and require control over individual electrons. Defects that are of little concern in room temperature devices become problematic at a low temperature and low electron density. For example, charges that are trapped in an insulating layer on the silicon can lead to localization of electrons in unintentional (and unwanted) parasitic quantum dots. In addition, various high-energy processing techniques, including ion implantation and plasma-based processes, are ubiquitous in modern integrated-circuit fabrication and can lead to trapped charges in the gate oxide and other dielectric layers. Nordberg and colleagues have investigated how various processing steps affect the performance of devices that are intended to manipulate one or a few electrons at millikelvin temperatures[65]. Through a combination of judicious choices of processing techniques and annealing, they have demonstrated working quantum devices, with few parasitic quantum dots.

Using heterostructures of Si and SiGe is one possible solution to meeting some of the challenges of defects. Instead of the electrons inhabiting an interface between crystalline silicon and an amorphous oxide, the electrons are held at the interface between two crystalline semiconductors. Electrons with very high mobility (indicative of a low density of defects and trapped charges) have been observed in these structures[66], which is a promising development; however, using these heterostructures is a step farther away from standard silicon processing. For example, maintaining the quality of the heterostructure interface imposes severe restrictions on the subsequent processes that can be used to define the devices.

Issues of defects, trapped charges and extra spins can ultimately be traced back to a modern silicon integrated circuit being made of many layers of different materials. An alternative approach is to use only crystalline silicon. At a low temperature, silicon becomes an insulator, except where it is degenerately doped. Simmons and co-workers have demonstrated how to control the incorporation of phosphorus at the atomic scale, allowing degenerate doping of local regions of the silicon with true atomic-scale resolution[67].

The successful development of processes for the fabrication of quantum devices in large numbers will open the door to the scaling of quantum processors. Questions of device uniformity, reproducibility and appropriate architecture will become important. For example, many designs require multiple classical control gates for each qubit, even though triple quantum-dot structures can be defined with only two top gates[68]. As noted earlier, mobile electrons in silicon lose their spin coherence much more rapidly than tightly bound spins; however, if the spins are to be localized, a high level of device reproducibility will be required to keep the number of control lines manageable[69].

Future directions

The approaches based on donor atoms and quantum dots each have their own advantages in terms of coherence times, control and read-out methods, and it is possible that both approaches could be combined in a future quantum processor. Indeed, this has already been shown to some extent in the single-shot measurement of a donor electron spin, by using controlled tunnelling onto a single-electron transistor (SET) quantum dot[42]. In both cases, it is clear that the single-qubit properties are excellent; they are sufficient for fault-tolerant quantum computing and are far ahead of other solid-state rival systems in many cases. A key goal of future work in silicon-based quantum computing is therefore to identify the best way in which to couple multiple, spatially separated qubits.

Spins on neighbouring qubits can interact through dipolar or exchange coupling. The exchange coupling between donors or quantum dots could be tuned using intermediate gates[70,71], although in the case of donors this would require ultra-precise positioning of the donors with respect to each other and the control gates. Dipolar coupling is typically a less attractive option than exchange coupling because of its weaker strength (leading to a slower gate); however, the long coherence times in silicon may allow it to be used. Interactions can be 'always on' (and still allow universal quantum computing) or turned on or off by moving the qubit into a nuclear-spin degree of freedom that shows much weaker dipole coupling[51]. An alternative route would be to dispense with the through-space coupling of spins and use the portable nature of the electron to move the spin qubit between different quantum dot or donor sites[25,72]. These different coupling methods must be assessed with respect to gate speed and fidelity, in addition to how well they lend themselves to scaling up based on the opportunities provided by silicon-based nanofabrication.

Silicon has been the material of choice for hosting the remarkable developments in information processing of the past 50 years, and it shows every promise of surviving the transition from the classical computer to the quantum computer. ■

1. Lansbergen, G. P. *et al*. Gate-induced quantum-confinement transition of a single dopant atom in a silicon FinFET. *Nature Phys.* **4,** 656–661 (2008).
2. Deutsch, D. Quantum theory, the Church–Turing principle and the universal quantum computer. *Proc. R. Soc. Lond. A* **400,** 97–117 (1985).
3. Ladd, T. D. *et al*. Quantum computers. *Nature* **464,** 45–53 (2010).
4. Hayashi, T., Fujisawa, T., Cheong, H. D., Jeong, Y. H. & Hirayama, Y. Coherent manipulation of electronic states in a double quantum dot. *Phys. Rev. Lett.* **91,** 226804 (2003).
5. Tyryshkin, A. M. *et al*. Electron spin coherence exceeding seconds in high purity silicon. *Nature Mater.* (in the press); preprint at ⟨http://arxiv.org/abs/1105.3772v1⟩ (2011).
This article reports the longest coherence time of any electron spin in the solid state; this was shown by donor electrons in silicon, with a T_2 of more than 10 s.
6. Steane, A. M. Efficient fault-tolerant quantum computing. *Nature* **399,** 124–126 (1999).
7. Andreas, B. *et al*. Determination of the Avogadro constant by counting the atoms in a ^{28}Si crystal. *Phys. Rev. Lett.* **106,** 030801 (2011).
8. Feher, G. Electron spin resonance experiments on donors in silicon. *Phys. Rev.* **114,** 1219–1244 (1959).
This article reports on a seminal study that investigated a wide range of spin properties of donors in silicon.
9. Gordon, J. & Bowers, K. Microwave spin echoes from donor electrons in silicon. *Phys. Rev. Lett.* **1,** 368–370 (1958).
10. Witzel, W. M. & Das Sarma, S. Quantum theory for electron spin decoherence induced by nuclear spin dynamics in semiconductor quantum computer architectures: spectral diffusion of localized electron spins in the nuclear solid-state environment. *Phys. Rev. B* **74,** 035322 (2006).
11. Abe, E. *et al*. Electron spin coherence of phosphorus donors in silicon: effect of environmental nuclei. *Phys. Rev. B* **82,** 121201 (2010).
12. Tyryshkin, A. M., Lyon, S. A., Astashkin, A. V. & Raitsimring, A. M. Electron spin relaxation times of phosphorus donors in silicon. *Phys. Rev. B* **68,** 193207 (2003).
13. Schenkel, T. *et al.* Electrical activation and electron spin coherence of ultra-low dose antimony implants in silicon. *Appl. Phys. Lett.* **88,** 112101 (2005).
14. Tyryshkin, A. M., Lyon, S. A., Jantsch, W. & Schäffler, F. Spin manipulation of free two-dimensional electrons in Si/SiGe quantum wells. *Phys. Rev. Lett.* **94,** 126802 (2005).
15. Appelbaum, I., Huang, B. & Monsma, D. J. Electronic measurement and control of spin transport in silicon. *Nature* **447,** 295–298 (2007).
16. Matsunami, J., Ooya, M. & Okamoto, T. Electrically detected electron spin resonance in a high-mobility silicon quantum well. *Phys. Rev. Lett.* **97,** 066602 (2006).
17. Wilamowski, Z., Malissa, H., Schaeffler, F. & Jantsch, W. *g*-Factor tuning and manipulation of spins by an electric current. *Phys. Rev. Lett.* **98,** 187203 (2007).
18. Shankar, S., Tyryshkin, A. M., He, J. & Lyon, S. A. Spin relaxation and coherence times for electrons at the Si/SiO_2 interface. *Phys. Rev. B* **82,** 195323 (2010).
19. Xiao, M., House, M. G. & Jiang, H. W. Measurement of the spin relaxation time of single electrons in a silicon metal-oxide-semiconductor-based quantum dot. *Phys. Rev. Lett.* **104,** 096801 (2010).

20. Hayes, R. R. *et al*. Lifetime measurements (T_1) of electron spins in Si/SiGe quantum dots. Preprint at ⟨http://arxiv.org/abs/0908.0173⟩ (2009).
References 19 and 20 report measurements of T_1 in spin qubits in Si/SiO_2 and Si/SiGe quantum dots.
21. Simmons, C. B. *et al*. Tunable spin loading and T_1 of a silicon spin qubit measured by single-shot readout. *Phys. Rev. Lett.* **106,** 156804 (2011).
This article was the first report of a single-shot spin read-out of a quantum-dot spin qubit in silicon, which was carried out in Si/SiGe.
22. Petta, J. R. *et al*. Coherent manipulation of coupled electron spins in semiconductor quantum dots. *Science* **309,** 2180–2184 (2005).
23. Schweiger, A. & Jeschke, G. Principles of pulse electron paramagnetic resonance (Oxford Univ. Press, 2001).
24. Viola, L., Knill, E. & Lloyd, S. Dynamical decoupling of open quantum systems. *Phys. Rev. Lett*. **82,** 2417–2421 (1999).
25. Morton, J. J. L. A silicon-based cluster state quantum computer. Preprint at ⟨http://arxiv.org/abs/0905.4008⟩ (2010).
26. Pioro-Ladrière, M. *et al*. Electrically driven single-electron spin resonance in a slanting Zeeman field. *Nature Phys*. **4,** 776–779 (2008).
27. Bradbury, F. R. *et al*. Stark tuning of donor electron spins in silicon. *Phys. Rev. Lett.* **97,** 176404 (2006).
28. Dreher, L. *et al*. Electroelastic hyperfine tuning of phosphorus donors in silicon. *Phys. Rev. Lett.* **106,** 037601 (2011).
29. Rugar, D., Budakian, R., Mamin, H. J. & Chui, B. W. Single spin detection by magnetic resonance force microscopy. *Nature* **430,** 329–332 (2004).
30. Schmidt, J. & Solomon, I. Modulation of the photoconductivity in silicon at low temperatures by electron magnetic resonance of shallow impurities. *C. R. Acad. Sci. III* **263,** 169–172 (1966).
31. McCamey, D. R. *et al*. Electrically detected magnetic resonance in ion-implanted Si:P nanostructures. *Appl. Phys. Lett.* **89,** 182115 (2006).
32. Stegner, A. R. *et al*. Electrical detection of coherent ^{31}P spin quantum states. *Nature Phys.* **2,** 835–838 (2006).
33. Morley, G. W. *et al*. Long-lived spin coherence in silicon with an electrical spin trap readout. *Phys. Rev. Lett.* **101,** 207602 (2008).
34. Boehme, C. & Lips, K. Spin-dependent recombination — an electronic readout mechanism for solid state quantum computers. *Phys. Status Solidi B* **233,** 427 (2002).
35. Lo, C. C., Bokor, J., Schenkel, T., Tyryshkin, A. M. & Lyon, S. A. Spin-dependent scattering off neutral antimony donors in ^{28}Si field-effect transistors. *Appl. Phys. Lett.* **91,** 242106 (2007).
36. Willems van Beveren, L. H. *et al*. Broadband electrically detected magnetic resonance of phosphorus donors in a silicon field-effect transistor. *Appl. Phys. Lett.* **93,** 072102 (2008).
37. Lo, C. C. *et al*. Electrically detected magnetic resonance of neutral donors interacting with a two-dimensional electron gas. *Phys. Rev. Lett*. **106,** 207601 (2011).
38. Xiao, M., Martin, I., Yablonovitch, E. & Jiang, H. W. Electrical detection of the spin resonance of a single electron in a silicon field-effect transistor. *Nature* **430,** 435–439 (2004).
39. Steger, M. *et al*. Optically detected NMR of optically hyperpolarized ^{31}P neutral donors in ^{28}Si. *J. Appl. Phys*. **109,** 102411 (2011).
40. Elzerman, J. M. *et al*. Single-shot read-out of an individual electron spin in a quantum dot. *Nature* **430,** 431–435 (2004).
41. Morello, A. *et al*. Architecture for high-sensitivity single-shot readout and control of the electron spin of individual donors in silicon. *Phys. Rev. B* **80,** 081307 (2009).
42. Morello, A. *et al*. Single-shot readout of an electron spin in silicon. *Nature* **467,** 687–691 (2010).
This article was the first report of a single-shot spin read-out of a single donor electron in silicon.
43. Tahan, C., Friesen, M. & Joynt, R. Decoherence of electron spin qubits in Si-based quantum computers. *Phys. Rev. B* **66,** 035314 (2002).
44. Levy, J. Universal quantum computation with spin-1/2 pairs and Heisenberg exchange. *Phys. Rev. Lett.* **89,** 147902 (2002).
45. Shaji, N. *et al*. Spin blockade and lifetime-enhanced transport in a few-electron Si/SiGe double quantum dot. *Nature Phys*. **4,** 540–544 (2008).
46. Liu, H. W. *et al*. Pauli-spin-blockade transport through a silicon double quantum dot. *Phys. Rev. B* **77,** 073310 (2008).
47. Lansbergen, G. P. *et al*. Lifetime-enhanced transport in silicon due to spin and valley blockade. *Phys. Rev. Lett.* **107,** 136602 (2011).
48. Morton, J. J. & Lovett, B. W. Hybrid solid-state qubits: the powerful role of electron spins. *Annu. Rev. Condens. Matter Phys.* **2,** 189–212 (2011).
49. Ladd, T. D., Maryenko, D., Yamamoto, Y., Abe, E. & Itoh, K. M. Coherence time of decoupled nuclear spins in silicon. *Phys. Rev. B* **71,** 014401 (2005).
50. Witzel, W. M. & Das Sarma, S. Nuclear spins as quantum memory in semiconductor nanostructures. *Phys. Rev. B* **76,** 045218 (2007).
51. Morton, J. J. L. *et al*. Solid state quantum memory using the ^{31}P nuclear spin. *Nature* **455,** 1085–1088 (2008).
This article reports the coherent transfer of quantum information between a donor electron spin and a coupled ^{31}P nuclear spin, yielding a nuclear T_2 of more than 1 s.
52. George, R. E. *et al*. Electron spin coherence and electron nuclear double resonance of Bi donors in natural Si. *Phys. Rev. Lett.* **105,** 067601 (2010).
53. McCamey, D. R., van Tol, J., Morley, G. W. & Boehme, C. Fast nuclear spin hyperpolarization of phosphorus in silicon. *Phys. Rev. Lett*. **102,** 027601 (2009).
54. Yang, A. *et al*. Simultaneous subsecond hyperpolarization of the nuclear and electron spins of phosphorus in silicon by optical pumping of exciton transitions. *Phys. Rev. Lett.* **102,** 1–4 (2009).
55. Simmons, S. *et al*. Entanglement in a solid-state spin ensemble. *Nature* **470,** 69–72 (2011).
56. Sarovar, M., Young, K. C., Schenkel, T. & Whaley, K. B. Quantum nondemolition measurements of single donor spins in semiconductors. *Phys. Rev. B* **78,** 245302 (2008).
57. McCamey, D. R., van Tol, J., Morley, G. W. & Boehme, C. Electronic spin storage in an electrically readable nuclear spin memory with a lifetime >100 seconds. *Science* **330,** 1652–1656 (2010).
This study demonstrated electrical read-out of both ^{31}P and ^{29}Si nuclear spin states in silicon, using the hyperfine interaction with donor electrons.
58. Neeley, M. *et al*. Generation of three-qubit entangled states using superconducting phase qubits. *Nature* **467,** 570–573 (2010).
59. DiCarlo, L. *et al*. Preparation and measurement of three-qubit entanglement in a superconducting circuit. *Nature* **467,** 574–578 (2010).
60. Wu, H. *et al*. Storage of multiple coherent microwave excitations in an electron spin ensemble. *Phys. Rev. Lett.* **105,** 140503 (2010).
61. Schuster, D. I. *et al*. High cooperativity coupling of electron-spin ensembles to superconducting cavities. *Phys. Rev. Lett.* **105,** 140501 (2010).
62. Kubo, Y. *et al*. Strong coupling of a spin ensemble to a superconducting resonator. *Phys. Rev. Lett.* **105,** 140502 (2010).
63. Zhu, X. *et al*. Coherent coupling of a superconducting flux qubit to an electron spin ensemble in diamond. *Nature* **478,** 221–224 (2011).
64. Politi, A., Cryan, M. J., Rarity, J. G., Yu, S. & O'Brien, J. L. Silica-on-silicon waveguide quantum circuits. *Science* **320,** 646–649 (2008).
65. Nordberg, E. P. *et al*. Enhancement-mode double-top-gated metal-oxide-semiconductor nanostructures with tunable lateral geometry. *Phys. Rev. B* **80,** 115331 (2009).
66. Lu, T. M., Tsui, D. C., Lee, C.-H. & Liu, C. W. Observation of two-dimensional electron gas in a Si quantum well with mobility of 1.6×10^6 cm^2/Vs. *Appl. Phys. Lett.* **94,** 182102 (2009).
67. Schofield, S. *et al*. Atomically precise placement of single dopants in Si. *Phys. Rev. Lett*. **91,** 136104 (2003).
This article describes the atomically precise placement of phosphorus donors in silicon, using the tip of a scanning tunnelling microscope.
68. Pierre, M. *et al*. Compact silicon double and triple dots realized with only two gates. *Appl. Phys. Lett*. **95,** 242107 (2009).
69. Oskin, M., Chong, F. T., Chuang, I. L. & Kubiatowicz, J. Building quantum wires: the long and the short of it. *Proc. Int. Symp. Comput. Architect.* 374–385 (ISCA, 2003).
70. Kane, B. E. A silicon-based nuclear spin quantum computer. *Nature* **393,** 133–137 (1998).
71. Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57,** 120–126 (1998).
References 70 and 71 were the first proposals of realistic architectures for implementing quantum information processing using donors and quantum dots in silicon.
72. Skinner, A., Davenport, M. & Kane, B. Hydrogenic spin quantum computing in silicon: a digital approach. *Phys. Rev. Lett*. **90,** 87901 (2003).
73. Haran, B. *et al*. 22 nm technology compatible fully functional 0.1 μm^2 6T-SRAM cell. *IEEE Electron Devices Meet.* 1–4 (IEEE, 2008).
74. Fuechsle, M. *et al*. Spectroscopy of few-electron single-crystal silicon quantum dots. *Nature Nanotechnol.* **5,** 502–505 (2010).
75. Morley, G.W. *et al*. Initializing, manipulating and storing quantum information with bismuth dopants in silicon. *Nature Mater.* **9,** 725–729 (2010).
76. Smelyanskiy, V. N., Petukhov, A. G. & Osipov, V. V. Quantum computing on long-lived donor states of Li in Si. *Phys. Rev. B* **72,** 081304 (2005).
77. Calderón, M., Koiller, B., Hu, X. & Das Sarma, S. Quantum control of donor electrons at the $Si–SiO_2$ interface. *Phys. Rev. Lett.* **96,** 096802 (2006).
78. Vrijen, R. *et al*. Electron-spin-resonance transistors for quantum computing in silicon-germanium heterostructures. *Phys. Rev. A* **62,** 012306 (2000).
79. Stoneham, A. M., Fisher, A. J. & Greenland, P. T. Optically driven silicon-based quantum gates with potential for high-temperature operation. *J. Phys. Condens. Matter* **15,** L447 (2003).
80. Schenkel, T. *et al*. Solid state quantum computer development in silicon with single ion implantation. *J. Appl. Phys.* **94,** 7017 (2003).
81. Andresen, S. *et al*. Charge state control and relaxation in an atomically doped silicon device. *Nano Lett.* **7,** 2000–2003 (2007).
82. Tan, K. Y. *et al*. Transport spectroscopy of single phosphorus donors in a silicon nanoscale transistor. *Nano Lett.* **10,** 11–15 (2010).
83. Sellier, H. *et al*. Transport spectroscopy of a single dopant in a gated silicon nanowire. *Phys. Rev. Lett.* **97,** 206805 (2006).
84. Lyding, J., Shen, T., Hubacek, J., Tucker, J. & Abeln, G. Nanoscale patterning and oxidation of H-passivated Si (100)−2×1 surfaces with an ultrahigh vacuum scanning tunneling microscope. *Appl. Phys. Lett.* **64,** 2010–2012 (1994).
85. Rokhinson, L. P., Guo, L. J., Chou, S. Y. & Tsui, D. C. Double-dot charge transport in Si single-electron/hole transistors. *Appl. Phys. Lett.* **76,** 1591 (2000).
86. Simmons, C. B. *et al*. Charge sensing and controllable tunnel coupling in a Si/SiGe double quantum dot. *Nano Lett.* **9,** 3234–3238 (2009).
87. Tracy, L. A. *et al*. Double quantum dot with tunable coupling in an enhancement-mode silicon metal-oxide semiconductor device with lateral geometry. *Appl. Phys. Lett.* **97,** 192110 (2010).
88. Sakr, M. R., Jiang, H. W., Yablonovitch, E. & Croke, E. T. Fabrication and characterization of electrostatic Si/SiGe quantum dots with an integrated read-out channel. *Appl. Phys. Lett*. **87,** 223104 (2005).
89. Berer, T. *et al*. Lateral quantum dots in Si/SiGe realized by a Schottky split-gate technique. *Appl. Phys. Lett*. **88,** 162112 (2006).

90. Angus, S. J., Ferguson, A. J., Dzurak, A. S. & Clark, R. G. Gate-defined quantum dots in intrinsic silicon. *Nano Lett*. **7,** 2051–2055 (2007).
91. Shin, Y.-S. *et al*. Aluminum oxide for an effective gate in Si/SiGe two-dimensional electron gas systems. *Semicond. Sci. Tech.* **26,** 055004 (2011).
92. Ruess, F. J. *et al*. Toward atomic-scale device fabrication in silicon using scanning probe microscopy. *Nano Lett.* **4,** 1969–1973 (2004).
93. Hu, Y. *et al*. A Ge/Si heterostructure nanowire-based double quantum dot with integrated charge sensor. *Nature Nanotechnol.* **2,** 622–625 (2007).
94. Zwanenburg, F. A., van Rijmenam, C. E. W. M., Fang, Y., Lieber, C. M. & Kouwenhoven, L. P. Spin states of the first four holes in a silicon nanowire quantum dot. *Nano Lett.* **9,** 1071–1079 (2009).
95. Simmons, C. B. *et al*. Single-electron quantum dot in Si/SiGe with integrated charge sensing. *Appl. Phys. Lett.* **91,** 213103 (2007).
96. Lim, W. H. *et al*. Observation of the single-electron regime in a highly tunable silicon quantum dot. *Appl. Phys. Lett*. **95,** 242102 (2009).
97. Weitz, P., Haug, R., von Klitzing, K. & Schäffler, F. Tilted magnetic field studies of spin- and valley-splittings in $Si/Si_{1-x}Ge_x$ heterostructures. *Surf. Sci.* **361–362,** 542–546 (1996).
98. Goswami, S. *et al*. Controllable valley splitting in silicon quantum devices. *Nature Phys*. **3,** 41–45 (2007).
99. Culcer, D., Cywinski, L., Li, Q. Z., Hu, X. & Das Sarma, S. Realizing singlet-triplet qubits in multivalley Si quantum dots. *Phys. Rev. B* **80,** 205302 (2009).
100. Friesen, M. & Coppersmith, S. N. Theory of valley-orbit coupling in a Si/SiGe quantum dot. *Phys. Rev. B* **81,** 115324 (2010).
101. Lai, N. S. *et al*. Pauli spin blockade in a highly tunable silicon double quantum dot. Preprint at http://arxiv.org/abs/1012.1410 (2010).

Acknowledgements We thank A. M. Tyryshkin for discussions. J.J.L.M. is supported by the Royal Society and St John's College, Oxford, and acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC) through the Centre for Advanced Electron Spin Resonance (EP/D048559/1) and the Japan Science and Technology Agency (JST)-EPSRC Cooperative Program (EP/H025952/1). D.R.M. is supported by an Australian Research Council Postdoctoral Fellowship (DP1093526). M.A.E. acknowledges support from the Army Research Office (ARO) (W911NF-08-1-0482). S.A.L. acknowledges support from the National Security Agency/Laboratory of Physical Sciences through Lawrence Berkeley National Laboratory (MOD 713106A), the National Science Foundation through the Princeton Materials Research Science and Engineering Center (DMR-0819860) and the ARO through Wisconsin. We apologize to those authors whose work could not be cited owing to space limitations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to J.J.L.M (john.morton@sjc.ox.ac.uk).

doi:10.1038/nature10682

Environmental effects of information and communications technologies

Eric Williams[1]

The digital revolution affects the environment on several levels. Most directly, information and communications technology (ICT) has environmental impacts through the manufacturing, operation and disposal of devices and network equipment, but it also provides ways to mitigate energy use, for example through smart buildings and teleworking. At a broader system level, ICTs influence economic growth and bring about technological and societal change. Managing the direct impacts of ICTs is more complex than just producing efficient devices, owing to the energetically expensive manufacturing process, and the increasing proliferation of devices needs to be taken into account.

One of the most striking aspects of information and communications technology (ICT) is the speed of its progress and adoption. Thirty years ago, information flows were mediated by postal deliveries, landline telephones and broadcast television, whereas now we access a globally interconnected world through a variety of devices from smart phones to large flat-screen displays. Technological progress in ICT is reflected in Moore's law, the observation that the number of transistors that can be packed into an integrated circuit doubles every 18 months[1]. Moore's law has become a self-fulfilling prophecy: the semiconductor industry actively aims to maintain Moore's rate of progress[2]. Although ICT currently relies on silicon-based integrated circuits, new technologies are on the horizon, including materials such as germanium and carbon, new architectures such as fin field-effect transistors (FinFETs), and new conceptual models such as quantum computing.

Few would dispute that ICTs are transforming societies and economies around the world. ICT is an example of a 'general-purpose technology', meaning that it interacts with and enhances other technologies[3]. Although the economic and social implications of ICTs are much discussed and analysed, the environmental implications receive much less attention. Yet ICTs interact fundamentally with environmental issues. To justify this assertion, first consider how previous technological revolutions, such as steam engines, the combustion engine and electricity, have fundamentally restructured human interactions with the environment. On the positive side, engines and electricity have greatly increased the efficiency of delivering energy services. At the same time, technology is a key element in an economic growth engine[4] that drives the increasing use of technology.

Consider, for example, the replacement of horses by automobiles in the twentieth century. Cars are much more efficient than horses in terms of environmental impact per distance travelled[5]. But their greater convenience and functionality, as well as their lower cost, mean that cars are used orders of magnitude more than horses ever were. Despite substantial improvements in efficiency and reductions in emissions during the twentieth century, the environmental challenges associated with automobiles remain unsolved. The key lesson here is that increasing the efficiency of a technology does not necessarily reduce its environmental impact.

ICTs interact with environmental issues at different system levels. Figure 1 depicts four types of interaction. The most direct and easily understood interaction is the physical layer, shown in the smallest circle in Fig. 1. At this level, ICT is physically embodied in an infrastructure and a set of devices whose manufacturing, operation and disposal have environmental impacts. At the next level, ICTs can be used to reduce environmental impacts with applications such as smart buildings, teleworking and optimized manufacturing[6–8]. Expanding the system boundary, ICTs contribute to economic growth[9] and shift consumption patterns[10,11]. At the broadest system level, ICTs are a key part of the info–nano–robotics–bio technological convergence that some believe will transform industry and society[12,13]. The capacity to comprehend and manage the different system levels decreases as the system boundary increases, as the higher levels are complex systems.

Here I discuss the various ways in which ICTs affect the environment. Society puts most emphasis on the direct impacts of ICT equipment, so these make up much of the article. I also discuss the implications of higher system levels, as in my opinion these are vastly more important than the direct impacts, despite receiving relatively little attention.

Assessing the environmental implications

Sustainability can be considered to be a societal reaction to human activities having global implications. The attempt to grapple with larger systems led to the development of methods to understand the systemic relationships between technology, environment and society. As discussed in the introduction, different levels of interaction involve different levels of complexity. New approaches are being developed to characterize different aspects of the problem. Not surprisingly, the degree of quantification and certainty decrease with increasing system levels.

Traditional methods of energy and environmental analysis can be used to characterize the first level of direct impacts. In addition, fresh insight can be obtained from newer approaches, such as material flow analysis (MFA) and life-cycle assessment (LCA). The first of these, MFA, is a general approach to measuring resource flows and emissions in industrial and natural systems[14], whereas LCA is a specialization and extension of MFA designed to quantify flows for the life cycle of a product or service[15,16]. LCA also includes impact assessment, which aims to characterize and assign trade-offs between different environmental impacts. Both approaches can be combined with forecasting approaches to estimate trends in macroscopic impacts[17], and both face the challenge of analysing complex and rapidly changing production chains and products.

The most popular form of LCA, the 'process-sum' method, builds a bottom-up model of materials flows facility by facility[15]. The complexity and proprietary information in ICT supply chains results in facility-level

[1]Golisano Institute of Sustainability, Rochester Institute of Technology, Rochester, New York 14623, USA.

data being unavailable for many processes, such as the purification of chemicals used to make semiconductors[18,19]. These data gaps mean that the process-sum method underestimates environmental impacts. Alternative LCA methods can address these data gaps, however. Economic input–output LCA (EIOLCA)[16,20], for example, is a holistic description of an economy based on a matrix of economic transactions between sectors[21]. It has the virtue of not excluding processes in the supply chain but has the disadvantage of aggregating multiple, sometimes diverse, processes into single sectors, which results in coarse-graining error[22]. Hybrid LCA aims to use both the process-sum and EIOLCA approaches together to minimize their weaknesses[18,19,23,24] but has yet to be widely adopted.

The environmental benefits of specific ICT applications such as teleworking, smart buildings and intelligent transport systems can be characterized by combining models of energy and materials use, technology and user behaviour. The main challenge is dealing with 'rebound effects', which occur when adopting a technology (such as a fuel-efficient vehicle) or practice (teleworking, say) indirectly induces additional impacts[25]. An economic rebound effect occurs when a technology change results in monetary savings that are then spent using that product more[26] or purchasing other environmentally intensive goods or services[25]. A second rebound effect occurs when saving time results in a behavioural change that induces further impacts, such as increased non-work driving by teleworkers[27].

The third and fourth system levels are more complex. Different disciplinary approaches can be applied to examine pieces of the system. The contributions of ICTs to economic growth are both direct, in terms of the economic output of ICT-related sectors, and indirect, by promoting growth in other sectors. Economists have explored the effects of ICT on economic growth by using neoclassical growth models (ref. 9). Further examples of disciplinary approaches are examining the effects of ICT on urban form through the lens of urban planning[28] and studying societal change through environmental sociology[29]. The different methodological approaches that could be applied are too numerous to recount here, but I make two high-level comments on the challenge of understanding the systemic implications of ICTs. First, an attempt to fuse disciplinary perspectives could bear fruit both in terms of methodological development and providing insight to manage the future. Second, given the complexity of the system, many important questions will be beyond the scope of quantitative or predictive modelling. Moving out through the system levels in Fig. 1, there is a transition from smaller subsystems that can be modelled with relative certainty to complex systems that are highly uncertain.

Most efforts to assess and manage the environmental implications of ICTs focus on direct impacts of ICT hardware. The primary issues of concern currently identified by society are the potential for exposure to hazardous materials and the use of energy, so I explore these in greater detail.

Exposure to hazardous materials

To review the relationship between ICT and exposure to hazards, first note that the impressive functionality of modern ICT is achieved by using a variety of exotic and highly refined materials both in products and as auxiliary materials in manufacturing. Given the wide range of materials used, it is not surprising that some are potentially hazardous.

It is important to distinguish between risk and hazard. Risk characterizes macroscopic health impacts, whereas hazard focuses on the potential for harm. Scientists and engineers generally gravitate towards a risk perspective, but the public sector, in particular non-governmental organizations, often take a hazard-based view that divides the world into acceptable and unacceptable materials (see, for example, ref. 30). Public policies to address hazardous materials in ICTs, including bans on landfill or materials such as lead-based solder, are based on a hazard perspective, as risk is little studied and poorly understood[31].

The primary concern for manufacturing is exposure to ancillary chemicals used in high-tech processing, in particular making semiconductors (see ref. 32 for recent results on cancer effects and ref. 33 for a review of earlier studies). In the operation of ICT devices, the main issue is exposure to brominated flame retardants (BFRs), which are added to casings and circuit boards in electronics, ostensibly to improve fire safety (see refs 34 and 35 for reviews of environmental issues for BFRs).

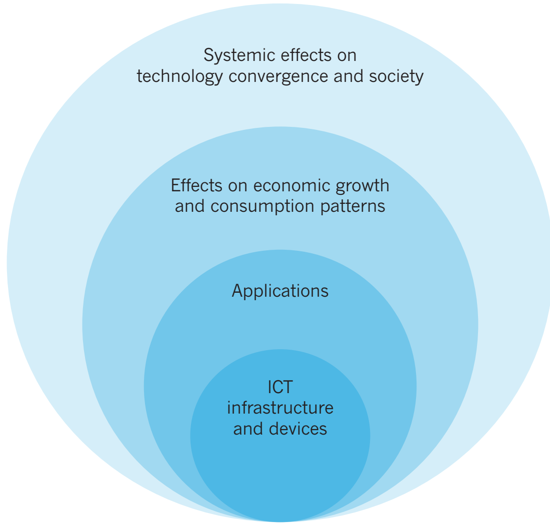


Figure 1 | Levels of system interactions between ICT and the environment. The inner circle shows direct impacts of ICT equipment and infrastructure. The second circle represents environmental applications of ICTs, such as teleworking. The third circle refers to effects on consumption caused by economic growth and changes in products. The outer circle represents larger societal and technological changes influenced by ICTs.

Potential exposure following the disposal of ICT devices has gained the most attention and is centred on three materials: metals, BFRs and compounds generated or used during recycling. An inventory of valuable and hazardous metals in a desktop computer system appears in Table 1. In addition to valuable metals such as copper, gold and silver, there are hazardous metals such as lead and cadmium. Hazardous materials are liberated or generated after disposal in three ways: leaching from landfills, incineration and recycling. Circuit boards and cathode-ray-tube monitors fail environmental regulatory tests for potential leaching from landfills, although there is scant evidence that leaching from sanitary landfills with leachate treatment systems poses a noticeable degree of risk[31]. Sometimes ICT devices are incinerated when inadvertently mixed into municipal waste streams, however, mobilizing hazardous metals and transforming BFRs into hazardous compounds such as brominated dioxins and furans. The degree to which combustion results in harmful emissions depends on the pollution controls at the incinerator.

Recycling is the third reason for post-disposal emissions. When recycling occurs in properly regulated facilities, efforts are made to ensure the safety of workers and the public alike. A great deal of ICT equipment is not recycled in proper facilities, however, but is processed by an informal (or backyard) industry in the developing world. With low labour cost and no environmental controls, recycling valuable metals from ICT devices in this way generates a profit, rather than incurring a net cost when faced with expensive labour and strict environmental controls. This economic situation drives the growth of an informal electronics industry in many parts of the developing world, such as China, India and Africa[36–38]. Copper is often recovered from wires by open burning of the insulation, which is usually made from polyvinyl chloride, so combustion releases dioxins, furans and other toxic chemicals. Gold in printed circuit boards is recovered by hydrometallurgical treatment using cyanide and acid without environmental controls. There is mounting evidence that informal recycling in the developing world is causing serious environmental pollution (see ref. 39 for a review).

Society's response to these hazardous concerns mainly takes the form of restricting the use of materials and establishing take-back

Table 1 | Quantities of valuable and hazardous metals in a desktop tower computer and cathode-ray-tube monitor[31]

Metal	Amount (g)
Aluminium	680–960
Antimony*	2.4–18.0
Arsenic*	0.06
Bismuth	0.23
Cadmium*	3.3
Chromium	0.05
Copper	1,370–2,640
Ferrite	480
Gold	0.39–0.67
Indium	0.04
Lead*	620–1,370
Nickel	4.5–30.0
Platinum	0.92
Steel	7,300–8,880
Silver	0.86–2.60
Tin	67
Zinc	21

*Most hazardous metals.

and recycling systems for electronics. The most prominent materials legislation is the European Restriction of Hazardous Substances Directive, which restricts the use of six substances — lead, mercury, cadmium, hexavalent chromium, polybrominated biphenyls (PBBs) and polybrominated diphenyl ether (PBDE) — in many electronics applications. Such restrictions and mandatory recycling presumably mitigate hazard, but they do not solve the most serious pollution issue, that of informal recycling in developing countries. One reason is that hazardous materials such as volatile organics and cyanide result from the recycling processes, rather than the toxic content of the products[31]. Second, difficulties in enforcement mean that exports of end-of-life electronics continue even when legislation is in place[40–42]. Third, global MFA forecasts indicate that by 2016–18 developing countries will generate more electronic waste than the developed world, and most of it will probably be recycled informally[17]. The heuristic goals that emerged from social processes (less toxic content and more recycling) need to be examined from a more systematic perspective.

Energy use

The effects of ICT hardware on energy and climate have come under increasing scrutiny. The life-cycle approach discussed above is crucial when examining the energy use of ICT because the production phase can be much more important for ICT than for other technologies, such as vehicles and buildings. For products with a plug or a fuel tank, the energy used during operation is far greater than that used during manufacturing. For example, 91% of the energy consumed by a typical home in Michigan is used while it is occupied, with just 9% used in materials and construction[43]; similarly, 88% of the energy that goes into a typical automobile relates to the fuel used to drive it[44]. The dominance of the use phase supports the traditional emphasis on improving operational energy efficiency.

For ICT devices, the energy used during manufacturing is a much bigger proportion. At the component level, LCA shows that at least 1.2 kg of fossil fuel is needed to manufacture a 2-g dynamic random access memory (DRAM) chip, a ratio of 600:1 compared with a 1:1 or 2:1 ratio for other manufactured goods[45]. The high energy use is due to the stringent purity standards required for semiconductor processing materials and environments. For a DRAM chip, at least 73% of the lifetime energy use goes into manufacturing, with just 27% used during operation. For computational logic chips, the roles are reversed, and operational energy (82%) far exceeds the energy used during manufacturing (18%)[46,47]. Note that both examples underestimate the contribution of the purification of chemicals and gases for semiconductor manufacturing, for which data are largely unavailable. Analysis of selected materials used in semiconductor manufacturing indicates that the energy used increases rapidly with higher purity[48].

For a typical laptop computer, hybrid LCA shows that 64% of the lifetime energy is used during manufacturing, with just 36% in operation[19]. This is partly because manufacturing computers is energy intensive, and partly because rapid obsolescence leads to computers being purchased more often than many other products with a plug. Figure 2 shows the ratio of manufacturing energy to operational life-cycle energy for ICT devices and other products.

The speed of technological progress in ICTs implies that environmental assessment must consider temporal change. One approach to assessing the effects of technological progress poses this question: how does the environmental impact per unit functionality change over time? The general answer is that it declines over time. For example, the electricity use per MHz in fabricating a desktop microprocessor has fallen from 0.028 kWh MHz^{-1} in 1995 to 0.001 kWh MHz^{-1} in 2006[49] as a result of improvements in manufacturing processes. The total energy use per functionality, including the supply chain for materials used in fabrication, has also decreased dramatically[47]. These results do not, however, include the increasingly stringent purity requirements for materials used in semiconductor manufacturing; higher purity could increase energy use[48].

It is not enough to consider environmental impact per functionality, however. As discussed in the introduction, the efficiency of automobiles, for example, has vastly improved over time but has coincided with rapidly increasing use of automobiles, which has brought new and as yet unsolved environmental challenges. One alternative to the functionality measure is to examine trends in environmental impact per typical product. As newer generations of products have additional functionality, it is instructive to compare the impact of succeeding generations. Here the ‘typical product’ measure parses out only the ‘per product’ piece of technological progress; the net impact from additional adoption is a separate factor. To complement the functionality measure (electricity per MHz) mentioned above, we can also track the electricity required to manufacture successive generations of typical desktop microprocessors. Measured in this way, the electricity used

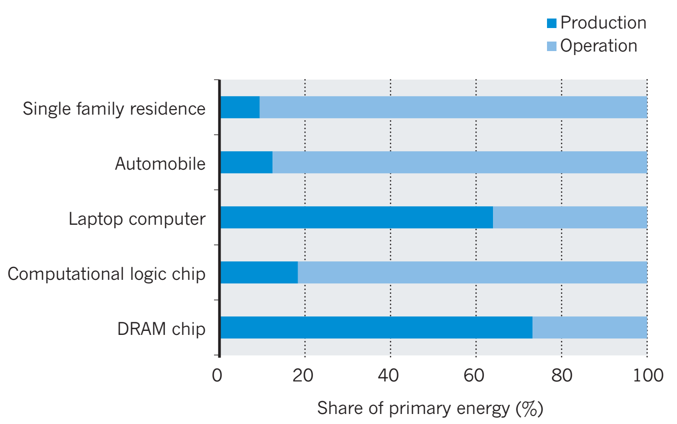


Figure 2 | Proportions of energy used in production and operation of various products. Figures for the DRAM chip, computational logic chip and laptop computer assume a three-year lifespan and home-use pattern of three hours per day, seven days per week. The energy used to run the logic chip, the automobile and the family home exceed the energy used in production, whereas the production energy is dominant for the DRAM chip and laptop computer.

per typical processor has not changed from 1995 to 2006[49]. Increased functionality — a faster processor — has cancelled out the efficiency gains per MHz. This example shows that progress in efficiency per functionality does not necessarily inform progress towards managing net environmental impacts.

The use of LCA often focuses on the product level. Although a product study provides information about the relative environmental impacts from different life-cycle stages, it does not address the question of macro-level impacts. To address net impacts, there is a stream of literature that estimates the total electricity use of ICT in different aggregations. For example, Kawamoto and collaborators[50] combined an inventory of computers, printers, copiers, faxes and networking devices in US offices with 'per device' data on electricity use. They found that in 2000, electricity use in computing and networking equipment in US offices was 74 TWh, or 2% of national electricity use[50]. A similar type of analysis covering 15 types of device, including televisions, computers, telephones and audio equipment, found that in 2002 consumer electronics in US homes consumed 147 TWh of electricity, or 4% of national electricity use[51]. Ignoring the different years of the studies, the two results combine to suggest that ICT devices in homes and offices constitute around 6% of US electricity use. The energy used to manufacture these devices has yet to be estimated, and there are no figures for temporal trends in energy overheads driven by changes in ownership and use.

Society's response to the energy used by ICTs has focused mainly on improving the efficiency of the operational phase of devices (the Energy Star standards set by the US Environmental Protection Agency for computers and displays, for example[52]). Manufacturers of components and devices have responded by making significant progress in improving operational energy efficiency. But it is not yet clear how well they are managing the net energy use over the devices' life cycle. In addition, there has been a continued proliferation of new ICT devices on the market, with the recent rise of smart phones, tablet computers and flat-screen televisions. The portfolio of ICT devices in use continues to increase, so reduced impacts for individual products may not lead to a reduced impact for the entire portfolio.

Managing future direct impacts

Nanoscale manipulation, which was pioneered in semiconductor manufacturing, has spun off into a nanotechnology industry. Given its potential use in everyday products such as textiles, paints and infrastructure, the scale of nanoparticle production could be orders of magnitude larger than for semiconductors alone. Semiconductor technology is also the basis for the manufacturing of photovoltaic modules, an area whose growth far outpaces the ICT industry. New materials and manufacturing processes are being developed for all of these applications. What are the strategic issues for assessing and managing the direct environmental impacts of these new technologies?

The pursuit of increased functionality is driving the use of more exotic materials, so concerns over potentially hazardous exposure will presumably continue. New issues are emerging, such as the effects of nanoparticles on health and ecosystems[53]. The public perception of hazard tends to dominate society's response, but this trend needs to be balanced with more science-based work that has a risk perspective.

Quiescent since the 1970s, concerns over the scarcity of resources have recently re-emerged. For electronics, the focus is on 'critical metals' such as tantalum, indium and ruthenium. Critical metals have supply constraints resulting from limited reserves, geopolitical problems or difficulty in recycling[54].

Societal response to managing the environmental impacts of ICTs has focused on heuristics such as removing toxic materials, increasing recycling and improving energy efficiency. As discussed earlier, following these heuristics has led to progress but has not solved the environmental challenges. In the future, managing the environmental problems facing ICTs will require a broader focus that considers life cycles, growth and technological progress.

Indirect effects

The direct environmental effects of ICT devices and infrastructure are important, but they should not distract us from the profound systemic environmental implications of ICTs. Environmentally beneficial applications need to be developed and promoted. Progress has been made in areas such as optimized control of manufacturing processes, but most of this progress has been achieved through market mechanisms rather than deliberate environmental intent. Public investments made to understand and use ICTs for environmental objectives pale in comparison with those made in industrial-age technologies such as engines, buildings, energy and road infrastructures.

Going beyond applications, the interaction of ICTs with economic growth, technological progress and society must not be ignored. The history of the automobile teaches an enduring lesson: improving a technology does not necessarily result in reduced environmental impacts. In fact, in an economy based on continuous growth rooted in technological progress, the opposite can be true. Understanding the interaction of ICTs with economic and social systems presents significant and interdisciplinary methodological challenges. Grappling with such complexity is at the heart of modern society's emerging concern over sustainability. ■

1. Moore, G. E. Cramming more components onto integrated circuits. *Electronics* **38,** 114–117 (1965).
2. ITRS International Technology Working Group *International Technology Roadmap for Semiconductors 2000–2009* ⟨http://www.itrs.net⟩.
3. Helpman, E. *General Purpose Technologies and Economic Growth* (Massachusetts Institut. Technol. Press, 1998).
4. Ayres, R. U. & Warr, B. Accounting for growth: the role of physical work. *Struct. Change Econ. Dyn.* **16,** 181–209 (2005).
This paper constructs an economic growth model in which useful work (from an energy perspective) is a major contributor to growth.
5. Grubler, A. *Technology and Global Change* (Cambridge Univ. Press, 1998).
6. Meyers, R., Williams, E. & Matthews, H. Scoping the potential of monitoring and control technologies to reduce energy use in U.S. homes. *Energy Build.* **42,** 563–569 (2010).
7. Matthews, H. S. & Williams, E. Telework adoption and energy use in building and transport sectors in the US and Japan. *J. Infrastruct. Syst.* **11,** 21–30 (2005).
8. Worrell, E., Martin, N. & Price, L. *Energy Efficiency and Carbon Dioxide Emissions Reduction Opportunities in the U.S. Iron and Steel Sector Report No.* LBNL-41724 (Lawrence Berkeley National Laboratory, 1999).
9. Organization for Economic Cooperation and Development. *ICT and Economic Growth. Evidence from OECD Countries, Industries and Firms* ⟨http://www.labs-associados.org/docs/OCDE_TIC.PDF⟩ (OECD, 2003).
10. Williams, E. & Hatanaka, T. Sustainable consumption and the information technology revolution. *Proc. First Int. Workshop Sustainable Consumption* 69–75 (Soc. Non-traditional Technol., 2003).
11. Hilty, L. *Information Technology and Sustainability: Essays on the Relationship* (Books on Demand, 2008).
12. Allenby, B. & Rejeski, D. The industrial ecology of emerging technologies. *J. Indust. Ecol.* **12,** 267–269 (2008).
13. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology* (Viking, 2005).
14. Graedel, T. & Allenby, B. *Industrial Ecology and Sustainable Engineering* 3rd edn (Prentice-Hall, 2009).
15. Baumann, H. & Tillman, A. M. *The Hitch Hiker's Guide to LCA: An Orientation in Life Cycle Assessment Methodology and Applications* (Studentlitteratur, 2004).
16. Hendrickson, C., Lave, L. & Matthews, S. *Environmental Life Cycle Assessment of Goods and Services: An Input–Output Approach* (RFF Press, 2006).
17. Yu, J., Williams, E., Ju, M. & Yang, Y. Forecasting global generation of obsolete personal computers. *Environ. Sci. Technol.* **44,** 3232–3237 (2010).
This global forecast predicts that the developing world will dispose of more computers than the developed world from 2016–18 onwards.
18. Williams, E. Energy intensity of computer manufacturing: hybrid analysis combining process and economic input–output methods. *Environ. Sci. Technol.* **38,** 6166–6174 (2004).
This paper develops a hybrid LCA method to account for missing data and finds that the energy used during manufacturing a home desktop computer exceeds its lifetime operating energy.
19. Deng, L., Babbitt, C. & Williams, E. Economic-balance hybrid LCA extended with uncertainty analysis: case study of laptop computer. *J. Cleaner Prod.* **19,** 1198–1206 (2011).
20. Bullard, C. & Herendeen, R. The energy cost of goods and services. *Energy Policy* **55,** 268–277 (1975).
21. Leontief, W. Quantitative input and output relations in the economic systems of the United States. *Rev. Econ. Stat.* **18,** 105–125 (1936).
22. Williams, E., Weber, C. & Hawkins, T. Hybrid approach to managing uncertainty in life cycle inventories. *J. Indust. Ecol.* **15,** 928–944 (2009).
23. Bullard, C., Pennter, P. & Pilati, D. Net energy analysis: handbook for combining process and input–output analysis. *Resour. Energy* **1,** 267–313 (1978).
24. Engelenburg, W., Van Rossum, M., Blok, K. & Vringer, K. Calculating the energy

requirements of household purchases: a practical step by step method. *Energy Policy* **21,** 648–656 (1994).
25. Hertwich, E. G. Consumption and the rebound effect: an industrial ecology perspective. *J. Indust. Ecol.* **9,** 85–98 (2005).
26. Greening, L., Greene, D. & Difiglio, C. Energy efficiency and consumption — the rebound effect — a survey. *Energy Policy* **28,** 389–401 (2000).
27. Mokhtarian, P. A synthetic approach to estimating the impacts of telecommuting on travel. *Urban Stud.* **35,** 215–241 (1998).
28. Audirac, I. Information technology and urban form: challenges to smart growth. *Int. Region. Sci. Rev.* **28,** 119–145 (2005).
29. Mol, A. Environmental governance in the Information Age: the emergence of informational governance. *Environ. Plan. C: Gov. Policy* **24,** 497–514 (2006).
30. Greenpeace Research Laboratories. *Missed Call: iPhone's Hazardous Chemicals* ⟨http://www.greenpeace.org/raw/content/international/press/reports/iPhones-hazardous-chemicals.pdf⟩ (Greenpeace International, 2007).
31. Williams, E. *et al.* Environmental, social and economic implications of global reuse and recycling of personal computers. *Environ. Sci. Technol.* **42,** 6446–6454 (2008).
This paper examines the sustainability implications for end-of-life computers, such as informal recycling in the developing world and emissions from electronics in landfill sites.
32. Boice, J. *et al.* Cancer mortality among US workers employed in semiconductor wafer fabrication. *J. Occup. Environ. Med.* **52,** 1082–1097 (2010).
33. Williams, E. The environmental impacts of semiconductor fabrication. *Thin Solid Films* **461,** 2–6 (2004).
34. Wikoff, D. & Birnbaum, L. Human health effects of brominated flame retardants. *Handbook Environ. Chem.* **16,** 19–53 (2011).
35. Shaw, S. *et al.* Halogenated flame retardants: do the fire safety benefits justify the risks? *Rev. Environ. Health* **25,** 261–305 (2010).
36. Basel Action Network & Silicon Valley Toxics Coalition. *Exporting Harm: The High-Tech Trashing of Asia* (Basel Action Network & Silicon Valley Toxics Coalition, 2002).
This non-governmental report highlighted the problem of informal recycling in the developing world.
37. Basel Action Network. *The Digital Dump: Exporting Re-use and Abuse to Africa*. *Media Release Version* (Basel Action Network, 2005).
38. Toxics Link. *Scrapping the High-Tech Myth: Computer Waste in India* (Toxics Link, 2003).
39. Tsydenova, O. & Bengtsson, M. Chemical hazards associated with treatment of waste electrical and electronic equipment. *Waste Manage.* **31,** 45–58 (2011).
40. Chisholm, M. & Bu, K. China's e-waste capital chokes on old computers. *Reuters News Service* (11 June 2007).
41. Warren, P. Organised crime targets waste recycling ⟨http://www.guardian.co.uk/technology/2009/jul/08/recycling-electronic-waste-crime⟩ *The Guardian* (8 July 2009).
42. Nordbrand, S. *Out of control: E-waste trade flows from the EU to developing countries* (SwedWatch, 2009).
43. Keolian, G., Blanchard, S. & Reppe, P. Life cycle energy, costs and strategies for improving a single family house. *J. Indust. Ecol.* **4,** 135–157 (2000).
44. Green Design Institute, Carnegie Mellon University. Economic Input–Output Life Cycle Assessment ⟨http://www.eiolca.net⟩ (Green Design Institute, 2011).
45. Williams, E., Ayres, R. & Heller, M. The 1.7 kg microchip: energy and chemical use in the production of semiconductors. *Environ. Sci. Technol.* **36,** 5504–5510 (2002).
46. Boyd, S., Horvath, A. & Dornfeld, D. Life-cycle energy demand and global warming potential of computational logic. *Environ. Sci. Technol.* **43,** 7303–7309 (2009).
47. Boyd, S., Horvath, A. & Dornfeld, D. Life-cycle assessment of computational logic produced from 1995 through 2010. *Environ. Res. Lett.* **5,** 014011 (2010).
48. Williams, E., Krishnan, N. & Boyd, S. in *Thermodynamics and the Destruction of Resources* (eds Bakshi, B., Gutowski, T. & Sekulic, D.) 190–211 (Cambridge Univ. Press, 2011).
49. Deng, L. & Williams, E. Functionality versus "typical product" measures of energy efficiency: case study of semiconductor manufacturing. *J. Indust. Ecol.* **15,** 108–121 (2011).
This paper develops the metric of 'typical product' to track efficiency trends and contrasts it with the standard functionality measure for energy use in semiconductor manufacturing.
50. Kawamoto, K. *et al. Electricity Used by Office Equipment and Network Equipment in the U.S.: Detailed Report and Appendices Report No.* LBNL-45917 (Lawrence Berkeley National Laboratory, 2001).
This study inventories US office and network equipment and links it to device-level energy-use data to estimate its national energy use.
51. Roth, K. W. & Kurtis McKenney, K. *Energy Consumption by Consumer Electronics in U.S. Residences* ⟨http://www.ce.org/pdf/Energy%20Consumption%20by%20CE%20in%20U.S.%20Residences%20(January%202007).pdf⟩ (TIAX, 2007).
52. US Environmental Protection Agency. *Energy Star* ⟨http://www.energystar.gov/index.cfm?fuseaction=find_a_product.showProductGroup&pgw_code=CO⟩ (Environmental Protection Agency, 2011).
53. Ok, Z., Benneyan, J. & Isaacs, J. Nanotechnology environmental, health and safety issues: brief literature review since 2000. 1–15 *Proc. IEEE Int. Symp. Sustainable Syst. Technol.* (IEEE, 2009).
54. Buchert, M., Schuler, D. & Bleher, D. *Critical Metals for Sustainable Technologies and their Recycling Potential* (United Nations Environment Programme, 2009).
This article discusses the definition of criticality of metals, recounts the use of critical metals in different technologies, and surveys the status of recycling.

Acknowledgements This work was in supported in part by the US National Science Foundation via grant CBET-0731067 in the Environmental Sustainability program.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to the author (exwgis@rit.edu).

doi:10.1038/nature10574

Species-specific responses of Late Quaternary megafauna to climate and humans

Eline D. Lorenzen[1]*, David Nogués-Bravo[2]*, Ludovic Orlando[1]*, Jaco Weinstock[1]*, Jonas Binladen[1]*, Katharine A. Marske[2]*, Andrew Ugan[3,42,43], Michael K. Borregaard[2], M. Thomas P. Gilbert[1], Rasmus Nielsen[4,5], Simon Y. W. Ho[6], Ted Goebel[7], Kelly E. Graf[7], David Byers[8], Jesper T. Stenderup[1], Morten Rasmussen[1], Paula F. Campos[1], Jennifer A. Leonard[9,10], Klaus-Peter Koepfli[11,12], Duane Froese[13], Grant Zazula[14], Thomas W. Stafford Jr[1,15], Kim Aaris-Sørensen[1], Persaram Batra[16], Alan M. Haywood[17], Joy S. Singarayer[18], Paul J. Valdes[18], Gennady Boeskorov[19], James A. Burns[20,21], Sergey P. Davydov[22], James Haile[1], Dennis L. Jenkins[23], Pavel Kosintsev[24], Tatyana Kuznetsova[25], Xulong Lai[26], Larry D. Martin[27], H. Gregory McDonald[28], Dick Mol[29], Morten Meldgaard[1], Kasper Munch[30], Elisabeth Stephan[31], Mikhail Sablin[32], Robert S. Sommer[33], Taras Sipko[34], Eric Scott[35], Marc A. Suchard[36,37], Alexei Tikhonov[32], Rane Willerslev[38], Robert K. Wayne[11], Alan Cooper[39], Michael Hofreiter[40], Andrei Sher[34]‡, Beth Shapiro[41], Carsten Rahbek[2] & Eske Willerslev[1]

Despite decades of research, the roles of climate and humans in driving the dramatic extinctions of large-bodied mammals during the Late Quaternary period remain contentious. Here we use ancient DNA, species distribution models and the human fossil record to elucidate how climate and humans shaped the demographic history of woolly rhinoceros, woolly mammoth, wild horse, reindeer, bison and musk ox. We show that climate has been a major driver of population change over the past 50,000 years. However, each species responds differently to the effects of climatic shifts, habitat redistribution and human encroachment. Although climate change alone can explain the extinction of some species, such as Eurasian musk ox and woolly rhinoceros, a combination of climatic and anthropogenic effects appears to be responsible for the extinction of others, including Eurasian steppe bison and wild horse. We find no genetic signature or any distinctive range dynamics distinguishing extinct from surviving species, emphasizing the challenges associated with predicting future responses of extant mammals to climate and human-mediated habitat change.

Towards the end of the Late Quaternary, beginning around 50,000 years ago, Eurasia and North America lost approximately 36% and 72% of their large-bodied mammalian genera (megafauna), respectively[1]. The debate surrounding the potential causes of these extinctions has focused primarily on the relative roles of climate and humans[2–5]. In general, the proportion of species that went extinct was greatest on continents that experienced the most dramatic climatic changes[6], implying a major role of climate in species loss. However, the continental pattern of megafaunal extinctions in North America and Australia approximately coincides with the first appearance of humans, suggesting a potential anthropogenic contribution to species extinctions[3,5].

[1]Centre for GeoGenetics, University of Copenhagen, Øster Voldgade 5–7, DK-1350 Copenhagen K, Denmark. [2]Center for Macroecology, Evolution and Climate, Department of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark. [3]Smithsonian Tropical Research Institute, Tupper Building, 401 Balboa, Ancón, Punamá, República de Panamá. [4]Departments of Integrative Biology and Statistics, University of California, Berkeley, 4098 VLSB, Berkeley, California 94720, USA. [5]Department of Biology, University of Copenhagen, Ole Maaloes Vej 5, DK-2200, Denmark. [6]School of Biological Sciences, University of Sydney, New South Wales 2006, Australia. [7]Center for the Study of the First Americans, Department of Anthropology, Texas A&M University, College Station, Texas 77843, USA. [8]Department of Sociology and Anthropology, Missouri State University, 901 South National, Springfield, Missouri 65807, USA. [9]Department of Evolutionary Biology, Uppsala University, 75236 Uppsala, Sweden. [10]Conservation and Evolutionary Genetics Group, Estación Biológica de Doñana (EBD-CSIC), Avenida Américo Vespucio, 41092 Seville, Spain. [11]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California 90095, USA. [12]Laboratory of Genomic Diversity, National Cancer Institute, Building 560, Room 11-33, Frederick, Maryland 21702, USA. [13]Department of Earth and Atmospheric Sciences, University of Alberta, Edmonton, Alberta T6G 2E3, Canada. [14]Government of Yukon, Department of Tourism and Culture, Yukon Palaeontology Program, PO Box 2703 L2A, Whitehorse, Yukon Territory Y1A 2C6, Canada. [15]Stafford Research Inc., 200 Acadia Avenue, Lafayette, Colorado 80026, USA. [16]Department of Earth and Environment, Mount Holyoke College, 50 College Street, South Hadley, Massachusetts 01075, USA. [17]School of Earth and Environment, University of Leeds, Woodhouse Lane, Leeds, West Yorkshire LS2 9JT, UK. [18]School of Geographical Sciences, University of Bristol, University Road, Bristol BS8 1SS, UK. [19]Diamond and Precious Metals Geology Institute, Siberian Branch of Russian Academy of Sciences, 39 Prospect Lenina, 677891 Yakutsk, Russia. [20]Royal Alberta Museum, Edmonton, Alberta T5N 0M6, Canada. [21]The Manitoba Museum, Winnipeg, Manitoba R3B 0N2, Canada. [22]North-East Science Station, Pacific Institute for Geography, Far East Branch of Russian Academy of Sciences, 2 Malinovy Yar Street, 678830 Chersky, Russia. [23]Museum of Natural and Cultural History, 1224 University of Oregon, Eugene, Oregon 97403-1224, USA. [24]Institute of Plant and Animal Ecology, Ural Branch of the Russian Academy of Sciences, 8 Marta Street, 202, 620144 Ekaterinburg, Russia. [25]Moscow State University, Vorob'evy Gory, 119899 Moscow, Russia. [26]State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Wuhan, Hubei 430074, China. [27]University of Kansas Museum of Natural History, University of Kansas, Lawrence, Kansas 66045, USA. [28]Park Museum Management Program, National Park Service, 1201 Oakridge Drive, Suite 150, Fort Collins, Colorado 80525, USA. [29]Natural History Museum, Rotterdam, c/o Gudumholm 41, 2133 HG Hoofddorp, Netherlands. [30]Bioinformatics Research Centre (BiRC), Aarhus University, C.F. Møllers Allé 8, DK-8000 Aarhus C, Denmark. [31]Regierungspräsidium Stuttgart, Landesamt für Denkmalpflege, Stromeyersdorfstrasse 3, D-78467 Konstanz, Germany. [32]Zoological Institute of Russian Academy of Sciences, Universitetskaya nab. 1, 199034 Saint-Petersburg, Russia. [33]Christian-Albrechts-University of Kiel, Institute for Nature and Resource Conservation, Department of Landscape Ecology, Olshausenstrasse 40, 24098 Kiel, Germany. [34]Institute of Ecology and Evolution, Russian Academy of Sciences, 33 Leninsky Prospect, 119071 Moscow, Russia. [35]San Bernardino County Museum, Division of Geological Sciences, 2024 Orange Tree Lane, Redlands, California 92374, USA. [36]Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California 90095, USA. [37]Department of Biostatistics, UCLA School of Public Health, University of California, Los Angeles, Los Angeles, California 90095, USA. [38]Museum of Cultural History, University of Oslo, St. Olavsgate 29, Postboks 6762 St. Olavsplass, 0130 Oslo, Norway. [39]Australian Centre for Ancient DNA, The University of Adelaide, South Australia 5005, Australia. [40]Department of Biology (Area 2), The University of York, Wentworth Way, Heslington, York YO10 5DD, UK. [41]Department of Biology, The Pennsylvania State University, 326 Mueller Laboratory, University Park, Pennsylvania 16802, USA. [42]Department of Anthropology, University of Utah, 271N1400E, Salt Lake City, Utah 84112-0060, USA. [43]Museo de Historia Natural de San Rafael, (5600) Parque Mariano Moveno, San Rafael, Mendoza, Argentina.
*These authors contributed equally to this work.
‡Deceased.

Demographic trajectories of different taxa vary widely and depend on the geographic scale and methodological approaches used[3,5,7]. For example, genetic diversity in bison[8,9], musk ox[10] and European cave bear[11] declines gradually from approximately 50,000–30,000 calendar years ago (kyr BP). In contrast, sudden losses of genetic diversity are observed in woolly mammoth[12,13] and cave lion[14] long before their extinction, followed by genetic stability until the extinction events. It remains unresolved whether the Late Quaternary extinctions were a cross-taxa response to widespread climatic or anthropogenic stressors, or were a species-specific response to one or both factors[15,16]. Additionally, it is unclear whether distinctive genetic signatures or geographical range-size dynamics characterize extinct or surviving species—questions of particular importance to the conservation of extant species.

To disentangle the processes underlying population dynamics and extinction, we investigate the demographic histories of six megafauna herbivores of the Late Quaternary: woolly rhinoceros (*Coelodonta antiquitatis*), woolly mammoth (*Mammuthus primigenius*), horse (wild *Equus ferus* and living domestic *Equus caballus*), reindeer/caribou (*Rangifer tarandus*), bison (*Bison priscus*/*Bison bison*) and musk ox (*Ovibos moschatus*). These taxa were characteristic of Late Quaternary Eurasia and/or North America and represent both extinct and extant species. Our analyses are based on 846 radiocarbon-dated mitochondrial DNA (mtDNA) control region sequences, 1,439 directly dated megafaunal remains and 6,291 radiocarbon determinations associated with Upper Palaeolithic human occupations in Eurasia. We reconstruct the demographic histories of the megafauna herbivores from ancient DNA data, model past species distributions and determine the geographical overlap between humans and megafauna over the past 50,000 years. We use these data to investigate how climate change and anthropogenic impacts affected species dynamics at continental and global scales, and contributed to the extinction of some species and the survival of others.

Responses differ among species and continents

The direct link between climate change, population size and species extinctions is difficult to document[10]. However, population size is probably controlled by the amount of available habitat and is indicated by the geographical range of a species[17,18]. We assessed the role of climate using species distribution models, dated megafauna fossil remains and palaeoclimatic data on temperature and precipitation. We estimated species range sizes at the time periods of 42, 30, 21 and 6 kyr BP as a proxy for habitat availability (Fig. 1 and Supplementary Information section 1). Range size dynamics were then compared with demographic histories inferred from ancient DNA using three distinct analyses (Supplementary Information section 3): (1) coalescent-based estimation of changes in effective population size through time (Bayesian skyride[19]), which allows detection of changes in global genetic diversity; (2) serial coalescent simulation followed by approximate Bayesian computation, which selects among different models describing continental population dynamics; and (3) isolation-by-distance analysis, which estimates potential population structure and connectivity within continents. If climate was a major factor driving species population sizes, we would expect expansion and contraction of a species' geographical range to mirror population increase and decline, respectively.

We find a positive correlation between changes in the size of available habitat and genetic diversity for the four species—horse, reindeer, bison and musk ox—for which we have range estimates spanning all four time-points (the correlation is not statistically significant for reindeer: $P = 0.101$) (Fig. 2 and Supplementary Information section 4). Hence, species distribution modelling based on fossil distributions and climate data are congruent with estimates of effective population size based on ancient DNA data, even in species with very different life-history traits. We conclude that climate has been a major driving force in megafauna population changes over the past 50,000 years. It is

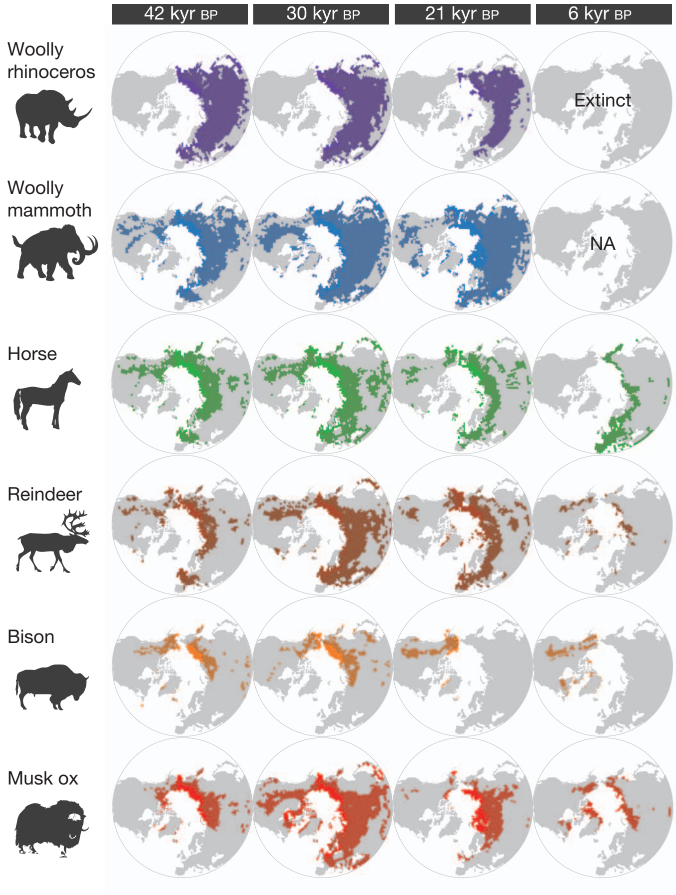


Figure 1 | Modelled potential ranges of megafauna species at 42, 30, 21 and 6 kyr BP. Ranges were modelled using the megafauna fossil record and palaeoclimatic data for temperature and precipitation; ice sheet extent was not included as a co-variable. Range measurements were restricted to the regions for which fossils were used to build the models, rather than all potentially suitable Holarctic area. NA, not available.

noteworthy that both estimated modelled ranges and genetic data are derived from a subset of the entire fossil record (Supplementary Information sections 1 and 3). Thus, changes in effective population size and range size might change with the addition of more data, especially from outside the geographical regions covered by the present study. However, we expect that the reported positive correlation will prevail when congruent data are compared.

The best-supported models of changes in effective population size in North America and Eurasia during periods of dramatic climatic change over the past 50,000 years are those in which populations increase in size (Fig. 3 and Supplementary Information section 3). This is true for all taxa except bison. However, the timing is not synchronous across populations. Specifically, we find highest support for population increase beginning approximately 34 kyr BP in Eurasian horse, reindeer and musk ox (Fig. 3a). Eurasian woolly mammoth and North American horse increase before the Last Glacial Maximum (LGM) approximately 26 kyr BP. Models of population increase in woolly rhinoceros and North American woolly mammoth fit equally well before and after the LGM, and North American reindeer populations increase later still. Only North American bison shows a population decline (Fig. 3b), the intensity of which probably swamps the signal of global population increase starting at approximately 35 kyr BP identified in the skyride plot (Fig. 2a).

These increases in effective population size probably reflect responses to climate change. By 34 kyr BP, the relatively warmer Marine Isotope Stage (MIS) 3 interstadial marked the transition to cold, arid full-glacial conditions of MIS 2, which began approximately

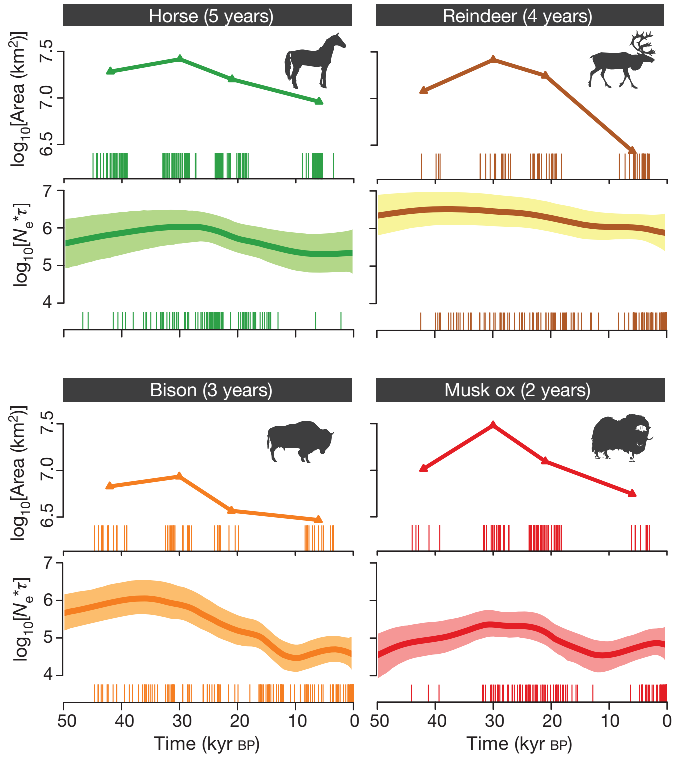


Figure 2 | Temporal changes in global genetic diversity and range size in horse, bison, reindeer and musk ox. The *x*-axis is in calendar years; the *y*-axis is the product of effective population size and generation time ($N_e\tau$). Generation times are given in parentheses. Comparable estimates of associated range sizes (square kilometres) are from Fig. 1. The temporal span of the radiocarbon-dated samples used in each approach is shown as vertical lines below each panel; each line represents one dated individual.

30 kyr BP[20,21]. Although the pre-LGM density of humans in Siberia remains uncertain, Pleistocene archaeological sites in the Siberian far north are scarce[22] and humans were presumably absent from North America before at least 15 kyr BP[23]. These point to climate, rather than humans, as the key driver of these species-specific and, in some cases, continent-specific demographic changes. This conclusion is supported by the significant correlations between modelled range sizes and effective population sizes (Fig. 2).

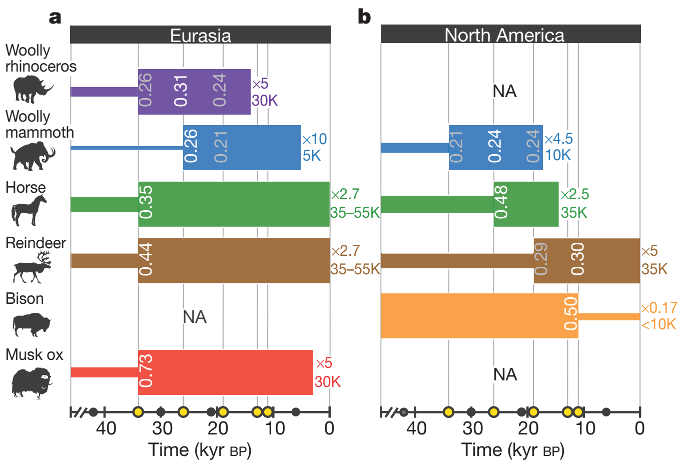


Figure 3 | Best-supported demographic models inferred by approximate Bayesian computation model-selection. **a**, Eurasia; **b**, North America. Grey dots on the time axis indicate periods with range size estimates. Yellow dots indicate the periods of demographic increase or decline, which were tested against each other in the approach. White values inside coloured bars reflect support for the best-supported model (for example, Eurasian woolly mammoth, increase at 26 kyr BP). The intensity of increase or decline (for example, ×10) and effective population size at the time of the youngest sample (for example, 5,000 individuals) are shown. We indicate in grey cases where multiple models received similar levels of support.

Modes of extinction

Both woolly rhinoceros and woolly mammoth suffered global extinctions during the Late Quaternary. Neither shows evidence of a decline in genetic diversity leading to their extinction at either continental or global scales (Supplementary Figs 3.2 and 3.6). However, the fossil records of the two species differ: woolly rhinoceros remains widely distributed across Eurasia until it disappears from the fossil record approximately 14 kyr BP (Supplementary Fig. 2.2), whereas the woolly mammoth range retreats northwards during its last millennia (Supplementary Figs 2.3 and 5.2c, d). We find increased isolation-by-distance preceding extinction (Supplementary Fig. 3.1 and Supplementary Information section 3), suggesting that populations of both species became increasingly fragmented, although the results are not statistically significant for woolly mammoth. The high and sustained levels of genetic diversity in these species might reflect the fixation of multiple distinct haplotypes in increasingly isolated and diminishing subpopulations. For woolly mammoth, this pattern is also supported by fossil evidence[24].

Our data suggest similar possibilities of increased isolation-by-distance before the extinctions of musk ox in Eurasia (approximately 2.5 kyr BP[25,26]) and of steppe bison in the north of the North American plains, which potentially survived until only a few hundred years ago[8] (Supplementary Fig. 3.1). Such fragmentation is commonly observed in wide-ranging species undergoing population decline, owing to populations aggregating in patches of high-quality habitat[27]. In contrast, we find low levels of isolation-by-distance in wild horse and in Eurasian and North American reindeer, suggesting these populations remained relatively panmictic over time.

Disentangling the roles of climate and humans

To evaluate the potential role of humans in the local and global megafauna extinctions, we measured the following: (1) the spatial overlap between the modelled range of each megafauna species and the Eurasian Palaeolithic archaeological record at 42, 30 and 21 kyr BP; (2) the presence of megafauna remains in Palaeolithic archaeological assemblages from Europe (48–18 kyr BP) and Siberia (41–12 kyr BP); and (3) variation in fossil abundance and the temporal and spatial distributions of known Palaeolithic archaeological sites and the Eurasian megafauna fossil record at 1,000-year intervals. For the last category, we added 1,557 indirectly dated megafaunal remains to the 1,439 directly dated specimens to increase sample sizes. Although associated with greater age-estimate uncertainties, the integrity of each of the indirectly dated samples was evaluated before inclusion following the guidelines listed in Supplementary Information section 5.

Woolly rhinoceros and Eurasian woolly mammoth experience a five- to tenfold increase in effective population size between 34 kyr BP and 19 kyr BP (Fig. 3), at least 10,000 years after first human contact as inferred from the overlap between estimated ranges and archaeological sites (Supplementary Figs 1.2 and 1.5). This result directly contradicts models of population collapse from human overkill (blitzkrieg)[2] or infectious diseases following the first human contact (hyperdisease)[28].

We find no evidence that Palaeolithic humans greatly impacted musk ox populations, in agreement with previous conclusions that humans were not responsible for the extinction of musk ox in Eurasia[10]. Musk ox remains are found in only 1% of European archaeological sites and 6% of Siberian sites, and do not overlap noticeably in range with Palaeolithic humans in either Europe or Siberia (Fig. 4). However, the decline in the potential range of musk ox by 60% between 21 and 6 kyr BP (Fig. 1), the increase in isolation-by-distance at 19 kyr BP (Supplementary Fig. 3.1 and Supplementary Table 3.3) and the positive correlation between climate-driven range size and genetic diversity (Fig. 2b) all point

towards climate as the main driver of musk ox population dynamics, including the decrease in genetic diversity after the LGM (Fig. 2a). The importance of climate is further supported by the physiology of musk ox, which might be a more sensitive indicator of environmental warming than the other species. Musk ox has extreme temperature sensitivity and is unable to tolerate high summer temperatures; the 10 °C summer isotherm approximates the southern limit of its present-day range[29].

We find little regional overlap between Palaeolithic humans and woolly rhinoceros in Siberia after the LGM (that is, after 20 kyr BP); the species is found in fewer than 11% of Siberian archaeological sites during this time (Fig. 4). This suggests that woolly rhinoceros was not a common prey species for humans, and that overhunting is an unlikely explanation for their extinction in Siberia. However, we note that geographical overlap existed between humans and woolly rhinoceros in Europe during the two millennia preceding extinction (Fig. 4), and therefore cannot exclude the hypothesis that humans influenced the final collapse of the species in this region. The continued presence of woolly rhinoceros in the fossil record throughout Siberia and parts of Europe up until the species extinction event (Supplementary Fig. 2.2) suggests that the final collapse of the species was synchronous across its range.

The data from woolly mammoth are inconclusive about the causes of extinction. We find that the range of Eurasian woolly mammoth overlaps continuously with humans throughout the Palaeolithic (Fig. 4), in agreement with previous results based on a more limited data set[30]. Woolly mammoth remains are found in 40% and 35% of all European and Siberian Palaeolithic sites, respectively, and mammoth subsistence hunting by Clovis peoples in North America has been documented[31]. However, the prevalence of woolly mammoth in Siberian sites declines after the LGM (43% of sites before 19 kyr BP compared with 30% after; Fig. 4). This decline could indicate a northward range shift of woolly mammoth ahead of humans[30] (Fig. 5.2c, d), an increasing scarcity of woolly mammoths in southern Siberia or an increasing human preference for other prey species.

In wild horse, the large mid-Holocene range of over 9 million km^2 (Fig. 1 and Supplementary Table 1.3) suggests the potential for a large Eurasian population at this time, and is not consistent with climate driving the final disappearance of the species in the wild. Rather, the decline in genetic diversity observed after the LGM in horse and bison, and to a lesser degree in reindeer (Fig. 2), might reflect the impact of expanding human populations in Europe and Asia. The presence of the three species in the archaeological record suggests that their populations are more likely to have been influenced by humans. Bison and horse are the most common megafauna herbivores found in archaeological sites (Fig. 4), with horse present in 58% and 66% of European and Siberian sites, respectively. Furthermore, horse shows extensive geographical

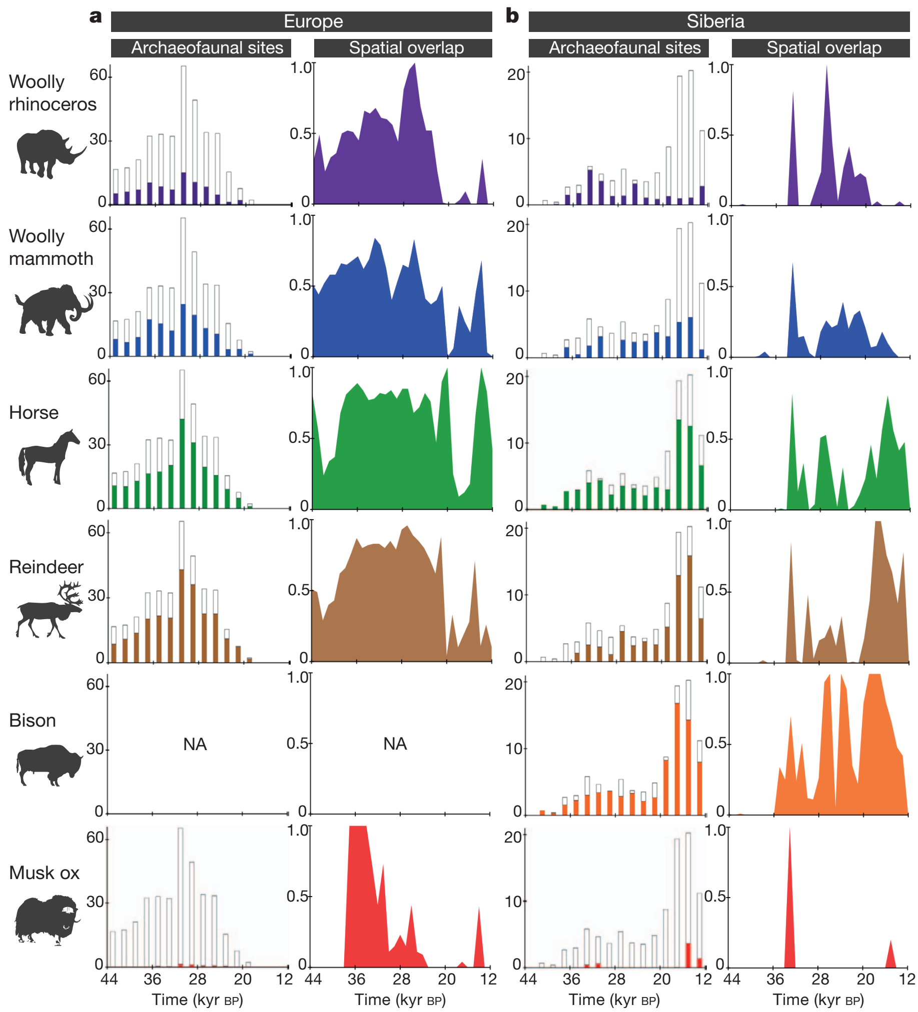


Figure 4 | Spatial and temporal association between megafauna and Upper Palaeolithic humans. **a**, Europe; **b**, Siberia. Column graphs represent known cultural occupations containing at least one of the six species, averaged over 2,000-year time bins; data span 48–18 kyr BP for Europe and 41–12 kyr BP for Siberia. Open bars indicate the number of archaeofaunal sites, filled bars represent the frequency of each species in the binned assemblages. Area graphs show the fraction of megafauna surface area shared with humans at 1,000-year intervals, calculated from the mean ±one standard deviation of latitude and longitude; data represented in Supplementary Fig. 5.2. Graphs use coordinates of data associated with both direct and indirect dates.

overlap with humans in both Europe and Siberia after the LGM, although large population sizes might have insulated horses to some extent from the effects of selective hunting by humans.

In bison, the pre-human decline in genetic diversity starting approximately 35 kyr BP and the strong correlation between range size and genetic diversity (Fig. 2) indicate climate as a main driver of demographic change. This conclusion is supported by the fivefold decline in effective population size (Fig. 3) and increased isolation-by-distance approximately 11 kyr BP in North America (Supplementary Fig. 3.1 and Supplementary Table 3.3). The timing of these demographic changes coincides with the pronounced climatic shifts associated with the Pleistocene/Holocene transition[32], although they also coincide with fossil evidence of growing populations of potential competitors such as *Alces* and *Cervus*[33]. The accelerated rate of decline in genetic diversity after approximately 16 kyr BP (Fig. 2) is coincident with the earliest known human expansion in the Americas[23], and the significant presence of bison in 77% of the Siberian archaeological assemblages points to their popularity as a prey species (Fig. 4).

Reindeer are the most abundant of the six taxa today. As with horse, they show continuous geographical overlap with Palaeolithic humans in Eurasia (Fig. 4). Reindeer are common in both European and Siberian Palaeolithic assemblages, are found in 67% of Siberian sites after the LGM and were an important prey species for humans in both Eurasia and North America[34]. Unlike bison and horse, the potential range of reindeer declines by 84% between 21 and 6 kyr BP (Fig. 1 and Supplementary Table 1.3). Despite the apparently detrimental influences of both humans and climate change, wild and domestic reindeer currently number in the millions across the Holarctic[35]. Although individual populations are affected by changing climate[36], the species is not currently under threat of extinction. The success of reindeer may be explained by high fecundity[37] and ecological flexibility[38]. In addition, continued low levels of isolation-by-distance suggest high mobility and near-panmixia of populations over millennia (Supplementary Fig. 3.1 and Supplementary Table 3.3).

Conclusions

We find that neither the effects of climate nor human occupation alone can explain the megafauna extinctions of the Late Quaternary. Rather, our results demonstrate that changes in megafauna abundance are idiosyncratic, with each species (and even continental populations within species) responding differently to the effects of climate change, habitat redistribution and human encroachment. Although reindeer remain relatively unaffected by any of these factors on a global scale, climate change alone explains the extinction of Eurasian musk ox and woolly rhinoceros, and a combination of climatic and anthropogenic effects appears to be responsible for the demise of wild horse and steppe bison. The causes underlying the extinction of woolly mammoth remain elusive.

We have shown that changes in habitat distribution and population size are intrinsically linked over evolutionary time, supporting the view that populations of many species will decline in the future owing to climate change and habitat loss. Intriguingly, however, we find no distinguishing characteristics in the rate or pattern of decline in those species that went extinct compared with those that have survived. Our study demonstrates the importance of incorporating lessons from the past into rational, data-driven strategies for the future to address our most pressing environmental challenges: the ongoing global mass-extinction of species and the impacts of global climate change and humans on the biodiversity that remains.

METHODS SUMMARY

Our data comprise 846 radiocarbon-dated ancient mitochondrial DNA sequences, 1,439 directly dated and 1,557 indirectly dated megafauna specimens, and 6,291 dated remains associated with Upper Palaeolithic humans in Eurasia. For population genetic analysis, we used the following: (1) the Bayesian skyride approach[20] to estimate the global demographic trajectory of each species over the past 50,000 years; (2) serial-coalescent simulations and the approximate Bayesian computation model-selection approach[39] to assess demographic change in Eurasia and in North America, and in the global data set; (3) isolation-by-distance to investigate changes in population structure over time in the two continental subpopulations. Palaeoclimatic estimates of precipitation and temperature were used to model the potential geographical range of each species at 42, 30, 21 and 6 kyr BP, using only contemporaneous radiocarbon-dated megafauna fossils (±3 kyr) for each period. Range measurements were restricted to Holarctic regions for which fossils were used to build the models. Using a Bayesian hierarchical modelling framework, these changes in range size were compared with changes in effective population size estimated from the Bayesian skyrides. To assess the spatial and temporal association between humans and megafauna, we (1) analysed variations in fossil abundance and spatial and temporal overlap between the human Upper Palaeolithic and megafauna fossil records in Europe and Siberia, (2) inferred the area of overlap between the human data from (1) and the megafauna ranges at 42, 30 and 21 kyr BP, and (3) assembled a list of the cultural occupations in Europe and Siberia with megafauna presence, to determine which taxa were directly associated with Palaeolithic humans. For details on methods see Supplementary Information.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 22 April; accepted 16 September 2011.
Published online 2 November 2011.

1. Barnosky, A. D., Koch, P. L., Feranec, R. S., Wing, S. L. & Shabel, A. B. Assessing the causes of Late Pleistocene extinctions on the continents. *Science* **306,** 70–75 (2004).
2. Martin, P. S. in *Quaternary Extinctions: A Prehistoric Revolution* (eds Martin, P. S. & Klein, R. G.) 364–403 (Univ. Arizona Press, 1984).
3. Alroy, J. A. multispecies overkill simulation of the end-Pleistocene megafaunal mass extinction. *Science* **292,** 1893–1896 (2001).
4. Stuart, A. J., Kosintsev, P. A., Higham, T. F. G. & Lister, A. M. Pleistocene to Holocene extinction dynamics in giant deer and woolly mammoth. *Nature* **431,** 684–689 (2004).
5. Koch, P. L. & Barnosky, A. D. Late Quaternary extinctions: state of the debate. *Annu. Rev. Ecol. Evol. Syst.* **37,** 215–250 (2006).
6. Nogués-Bravo, D., Ohlemüller, R., Batra, P. & Araújo, M. B. Climate predictors of Late Quaternary extinctions. *Evolution* **64,** 2442–2449 (2010).
7. Haile, J. *et al.* Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl Acad. Sci. USA* **106,** 22363–22368 (2009).
8. Shapiro, B. *et al.* Rise and fall of the Beringian steppe bison. *Science* **306,** 1561–1565 (2004).
9. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22,** 1185–1192 (2005).
10. Campos, P. F. *et al.* Ancient DNA analyses exclude humans as the driving force behind Late Pleistocene musk ox (*Ovibos moschatus*) population dynamics. *Proc. Natl Acad. Sci. USA* **107,** 5675–5680 (2010).
11. Stiller, M. *et al.* Withering away—25,000 years of genetic decline preceded cave bear extinction. *Mol. Biol. Evol.* **27,** 975–978 (2010).
12. Barnes, I. *et al.* Genetic structure and extinction of the woolly mammoth, *Mammuthus primigenius. Curr. Biol.* **17,** 1–4 (2007).
13. Debruyne, R. *et al.* Out of America: ancient DNA evidence for a new world origin of Late Quaternary woolly mammoths. *Curr. Biol.* **18,** 1320–1326 (2008).
14. Barnett, R., Yamaguchi, N., Barnes, I. & Cooper, A. The origin, current diversity, and future conservation of the modern lion (*Panthera leo*). *Proc. R. Soc. B* **273,** 2159–2168 (2006).
15. Guthrie, R. D. Rapid body size decline in Alaskan Pleistocene horses before extinction. *Nature* **426,** 169–171 (2003).
16. Grayson, D. K. Deciphering North American Pleistocene extinctions. *J. Anthropol. Res.* **63,** 185–214 (2007).
17. Andrewartha, H. G. & Birch, L. C. *The Distribution and Abundance of Animals* (Univ. Chicago Press, 1954).
18. Borregaard, M. K. & Rahbek, C. Causality in the relationship between geographic distribution and species abundance. *Q. Rev. Biol.* **85,** 3–25 (2010).
19. Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25,** 1459–1471 (2008).
20. Zazula, G. D. *et al.* Ice age steppe vegetation in east Beringia. *Nature* **423,** 603 (2003).
21. Zazula, G. D., Froese, D. G., Elias, S. A., Kuzmina, S. & Mathewes, R. W. Arctic ground squirrels of the mammoth-steppe: paleoecology of Late Pleistocene middens (~24 000–29 450 ^{14}C yr BP), Yukon Territory, Canada. *Quat. Sci. Rev.* **26,** 979–1003 (2007).
22. Pitulko, V. V. *et al.* The Yana RHS site: humans in the Arctic before the last glacial maximum. *Science* **303,** 52–56 (2004).
23. Goebel, T., Waters, M. R. & O'Rourke, D. H. The Late Pleistocene dispersal of modern humans in the Americas. *Science* **319,** 1497–1502 (2008).

24. Stuart, A. J., Sulerzhitsky, L. D., Orlova, L. A., Kuzmin, Y. V. & Lister, A. M. The latest woolly mammoths (*Mammuthus primigenius* Blumenbach) in Europe and Asia: a review of the current evidence. *Quat. Sci. Rev.* **21,** 1559–1569 (2002).
25. Vereshchagin, N. K. Prehistoric hunting and the extinction of Pleistocene mammals in the USSR. *Proc. Zool. Inst. Russ. Acad. Sci.* **69,** 200–232 (1971).
26. Kuznetsova, T. V., Sulerzhitsky, L. D., Siegert, C. & Schirrmeister, L. (2001) in *La Terra degli Elefanti* (eds Cavarretta, G. Giola, P., Mussi, M. & Palombo, M. R.) *The World of Elephants, Proc. 1st Int. Congr.* 289–292 (2001).
27. Wilson, R. J., Thomas, C. D., Fox, R., Roy, D. B. & Kunin, W. E. Spatial patterns in species distributions reveal biodiversity change. *Nature* **432,** 393–396 (2004).
28. MacPhee, R. D. E. & Marx, P. A. in *Natural Change and Human Impact in Madagascar* (eds Goodman, S. M. & Patterson, B. D.) 169–217 (Smithsonian Institution Press, 1997).
29. Tener, J. S. Muskoxen in Canada: a biological and taxonomic review. *Canadian Wildlife Service Monograph Series No. 2* (1965).
30. Ugan, A. & Byers, D. A global perspective on the spatiotemporal pattern of the Late Pleistocene human and woolly mammoth radiocarbon record. *Quaternary Int.* **191,** 69–81 (2008).
31. Surovell, T. A. & Waguespack, N. M. How many elephant kills are 14? Clovis mammoth and mastodon kills in context. *Quaternary Int.* **191,** 82–97 (2008).
32. Zielinski, G. A. & Mershon, G. R. Paleoenvironmental implications of the insoluble microparticle record in the GISP2 (Greenland) ice core during the rapidly changing climate of the Pleistocene–Holocene transition. *Geol. Soc. Am. Bull.* **109,** 547–559 (1997).
33. Guthrie, R. D. New carbon dates link climatic change with human colonization and Pleistocene extinctions. *Nature* **441,** 207–209 (2006).
34. Farnell, R. *et al.* Multidisciplinary investigations of alpine ice patches in southwest Yukon, Canada: paleoenvironmental and paleobiological investigations. *Arctic* **57,** 247–259 (2004).
35. Williams, T. M. & Heard, D. C. World status of wild *Rangifer tarandus* population. *Rangifer* **1** (special issue), 19–28 (1986).
36. Joly, K., Klein, D. R., Verbyla, D. L., Rupp, T. S. & Chapin, F. S. Linkages between large-scale climate patterns and the dynamics of Arctic caribou populations. *Ecography* **34,** 345–352 (2011).
37. Skogland, T. The effects of density-dependent resource limitation on the demography of wild reindeer. *J. Anim. Ecol.* **54,** 359–374 (1985).
38. Leader-Williams, N. *Reindeer on South Georgia* Ch. 1, 3–18 (Cambridge Univ. Press, 1988).
39. Beaumont, M. in *Simulations, Genetics and Human Prehistory* (eds Matsumura, S., Forster, P. & Renfrew, C.) 134–154 (McDonald Institute for Archaeological Research, 2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This paper is in memory of our friend and colleague Andrei Sher, who was a contributor to this study. Dr Sher died unexpectedly, but his major contributions to the field of Quaternary science will be remembered and appreciated for many years. We are grateful to A. Lister and T. Stuart for guidance and discussions. We thank T. B. Brandt, B. Hockett and A. Telka for laboratory help and samples, and L. M. R. Thrane for his work on the megafauna locality database. Data taken from the Stage 3 project were partly funded by grant F/757/A from the Leverhulme Trust, and a grant from the McDonald Grants and Awards Fund. B.S. was supported by NSF ARC-0909456. We acknowledge the Danish National Research Foundation, the Lundbeck Foundation, the Danish Council for Independent Research and the US National Science Foundation for financial support.

Author Contributions E.W. conceived and headed the overall project. C.R. headed the species distribution modelling and range measurements. E.D.L. and J.T.S. extracted, amplified and sequenced the reindeer DNA sequences. J.B. extracted, amplified and sequenced the woolly rhinoceros DNA sequences; M.H. generated part of the woolly rhinoceros data. J.W., K.-P.K., J.L. and R.K.W. generated the horse DNA sequences; A.C. generated part of the horse data. L.O., E.D.L. and B.S. analysed the genetic data, with input from R.N., K.M., M.A.S. and S.Y.W.H. Palaeoclimate simulations were provided by P.B., A.M.H, J.S.S. and P.J.V. The directly dated spatial latitudinal/longitudinal megafauna locality information was collected by E.D.L., K.A.M., D.N.-B., D.B. and A.U.; K.A.M. and D.N.-B. performed the species distribution modelling and range measurements. M.B. carried out the gene–climate correlation. A.U. and D.B. assembled the human Upper Palaeolithic sites from Eurasia. T.G. and K.E.G. assembled the archaeofaunal assemblages from Siberia. A.U. analysed the spatial overlap of humans and megafauna and the archaeofaunal assemblages. E.D.L., L.O., B.S., K.A.M., D.N.-B., M.K.B., A.U., T.G. and K.E.G. wrote the Supplementary Information. D.F., G.Z., T.W.S., K.A.-S., G.B., J.A.B., D.L.J., P.K., T.K., X.L., L.D.M., H.G.M., D.M., M.M., E.S., M.S., R.S.S., T.S., E.S., A.T., R.W. and A.C. provided the megafauna samples used for ancient DNA analysis. E.D.L. produced the figures. E.D.L, L.O. and E.W. wrote most of the manuscript, with input from B.S., M.H., D.N.-B., K.A.M., M.T.P.G., C.R., R.K.W, A.U. and the remaining authors.

Author Information Mitochondrial DNA sequences are deposited in GenBank under accession numbers JN570760–JN571033. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.W. (ewillerslev@snm.ku.dk).

METHODS

Data. Mitochondrial DNA sequences and accelerator mass spectrometry radiocarbon dates were collected from the past and present geographical ranges of six megafauna herbivores from Eurasia and North America: woolly rhinoceros (*Coelodonta antiquitatis*), woolly mammoth (*Mammuthus primigenius*), horse (wild *Equus ferus* and living domestic *Equus caballus*), reindeer/caribou (*Rangifer tarandus*), bison (*Bison priscus*/*Bison bison*) and musk ox (*Ovibos moschatus*) (Supplementary Fig. 2.1 and Supplementary Information sections 2 and 3). Our data comprise 846 radiocarbon-dated ancient mitochondrial DNA sequences (274 of which are new), 1,439 directly dated megafauna specimens (357 of which are new) and 6,291 dated remains associated with Upper Palaeolithic humans in Eurasia. In one analysis of the spatial and temporal association between humans and megafauna detailed below, we included an additional 1,557 indirectly dated megafaunal remains.

Species distribution modelling. We assessed changes in potential range size of each species over the past 50,000 years using 829 radiocarbon-dated megafauna fossils calibrated with the IntCal09 calibration curve[40] and palaeoclimatic estimates of precipitation and temperature[41]. Potential ranges were estimated for the four periods for which palaeoclimatic data are available, 42, 30, 21 and 6 kyr BP, using only contemporaneous fossils ($\pm$3 kyr) for each period (Supplementary Fig. 1.2). We compared temporal changes in potential range size (from species distribution models) and genetic diversity (from Bayesian skyrides[19]) during the past 50 kyr BP to assess the relation between these independent proxies of population size. If climate were a major driver of changes in population size, we would expect these two measures to be positively correlated. Estimating past ranges using species distribution models can be affected by an incomplete or biased fossil record as well as inaccuracies in the palaeoclimate simulations used in the models; uncertainties associated with these issues are depicted in our estimates of range size and how it correlates to genetic diversity (Supplementary Fig. 4.3). Range measurements were restricted to regions for which fossils were used to build the models, rather than all potentially suitable Holarctic areas. Fossil localities represent a subset, rather than an exhaustive search, of the literature available, and modelled ranges consequently represent a subset of the entire past distribution of the species. Too few fossils were available to estimate the potential ranges of woolly rhinoceros and woolly mammoth at 6 kyr BP, as the former was extinct and the latter was restricted to two island populations. Thus, too few periods with range estimates for these two species precluded statistical comparison with the genetic data, which spanned 50,000 years. For further details see Supplementary Information sections 1 and 4.

Ancient genetic analysis. We used three analytical approaches capable of incorporating serially sampled data to reconstruct the past population dynamics of each megafauna herbivore species. (1) The Bayesian skyride approach[19] estimates changes in genetic diversity through time as a proxy for effective population size, and was used to estimate the global demographic trajectory of each species. Because these data sets comprise samples from both a broad temporal and geographical extent, it is likely that they violate, at least during some of their evolutionary history, the assumption of panmixia made by the coalescent models currently implemented in BEAST[42]. However, the skyride makes the least stringent prior assumptions among these coalescent models, and therefore is the most likely to accommodate the temporal changes in structure that might characterize each of these species. (2) Serial-coalescent simulations and the approximate Bayesian computation model-selection approach[39] were used to test for demographic change in the continental subpopulations (Eurasia and North America) and in the global data set. Time points were chosen to represent midpoints between the four periods (42, 30, 21 and 6 kyr BP) for which we modelled potential megafauna ranges, and periods of dramatic climatic changes: the beginning (26 kyr BP) and end (19 kyr BP) of the LGM, the onset of the Younger Dryas (12.9 kyr BP) and the beginning of the Holocene (11 kyr BP). (3) Isolation-by-distance was used to test for changes in population structure over time in the continental subpopulations. Note that as with the species distribution models, the demographic events inferred from the ancient DNA data are conditional upon the samples included in the analysis. Hence, although we use the broad geographical terms of Eurasia and North America, the regions are limited to the localities covered by the sequenced samples (Supplementary Fig. 2.1). For further details on the genetics data see Supplementary Information section 2. For further details on the statistical analysis see Supplementary Information section 3.

Spatial association between megafauna and Palaeolithic humans. The presence of humans within the range of a species might directly or indirectly influence the capacity of the species to occupy that habitat. As a proxy for human impact, we assessed the spatial and temporal association between humans and megafauna using three approaches. (1) We compiled the human Upper Palaeolithic fossil record (50–12 kyr BP), including 6,291 radiocarbon determinations associated with human occupations in Europe and Siberia. We analysed variations in fossil abundance and spatial and temporal overlap at 1,000-year intervals between humans and the megafauna fossil record. To increase sample sizes for this particular analysis, we augmented the 1,439 directly dated megafauna specimens with an additional 1,557 indirectly dated megafaunal remains. Although associated with greater age-estimate uncertainties, the integrity of each indirectly dated sample was evaluated before inclusion following the guidelines listed in Supplementary Information section 5. (2) We inferred the area of overlap between the archaeological record from (1) and the megafauna ranges at 42, 30 and 21 kyr BP estimated using species distribution models. (3) We assembled a list of 380 cultural occupations in Europe (48–18 kyr BP) and 98 sites in Siberia (41–12 kyr BP) with megafauna presence, to determine which taxa were directly associated with Palaeolithic humans. For further details see Supplementary Information section 5.

40. Reimer, P. J. *et al.* IntCal09 and Marine09 radiocarbon age calibration curves, 0–50,000 years cal BP. *Radiocarbon* **51,** 1111–1150 (2009).
41. Nogués-Bravo, D., Rodríguez, J., Hortal, J., Batra, P. & Araújo, M. B. Climate change, humans, and the extinction of the woolly mammoth. *PLoS Biol.* **6,** e79 (2008).
42. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).

doi:10.1038/nature10572

Transgenerational epigenetic inheritance of longevity in *Caenorhabditis elegans*

Eric L. Greer[1,2], Travis J. Maures[1], Duygu Ucar[1], Anna G. Hauswirth[1], Elena Mancini[1], Jana P. Lim[1], Bérénice A. Benayoun[1], Yang Shi[2] & Anne Brunet[1]

Chromatin modifiers regulate lifespan in several organisms, raising the question of whether changes in chromatin states in the parental generation could be incompletely reprogrammed in the next generation and thereby affect the lifespan of descendants. The histone H3 lysine 4 trimethylation (H3K4me3) complex, composed of ASH-2, WDR-5 and the histone methyltransferase SET-2, regulates *Caenorhabditis elegans* lifespan. Here we show that deficiencies in the H3K4me3 chromatin modifiers ASH-2, WDR-5 or SET-2 in the parental generation extend the lifespan of descendants up until the third generation. The transgenerational inheritance of lifespan extension by members of the ASH-2 complex is dependent on the H3K4me3 demethylase RBR-2, and requires the presence of a functioning germline in the descendants. Transgenerational inheritance of lifespan is specific for the H3K4me3 methylation complex and is associated with epigenetic changes in gene expression. Thus, manipulation of specific chromatin modifiers only in parents can induce an epigenetic memory of longevity in descendants.

Transgenerational epigenetic inheritance has been described for some traits, including flower symmetry and colour in plants[1–3], progeny production in worms[4], heat stress response and eye colour in *Drosophila*[5–7], and coat colour in mammals[8–10]. However, the transgenerational epigenetic inheritance of longevity, and more generally of complex traits, is largely undefined. Chromatin modifiers have been shown to regulate longevity in several species[11–18], raising the possibility that chromatin changes in parents might not be entirely reset between generations and thereby also regulate longevity in descendants. Deficiencies in the H3K4me3 regulatory complex composed of ASH-2, WDR-5 and SET-2 extend lifespan in *C. elegans*[12]. We asked if perturbation of members of the H3K4me3 regulatory complex (ASH-2, WDR-5 and SET-2) only in the parental generation could regulate the lifespan of descendants in subsequent generations in *C. elegans*.

Transgenerational inheritance of longevity

We first focused on WDR-5, a conserved regulatory component of the ASH-2 complex[19] whose depletion decreases H3K4me3 levels[12,20–22] and extends lifespan in worms[12]. To test whether longevity could be inherited in a transgenerational epigenetic manner, we crossed wild-type (+/+) males with *wdr-5*(*ok1417*) mutant (*wdr-5*/*wdr-5*) hermaphrodites to generate F1 heterozygous hermaphrodites (Fig. 1a). These F1 heterozygous hermaphrodites were genotyped and then self-crossed to generate F2 hermaphrodites (wild type, heterozygous and homozygous at the *wdr-5* locus), which were genotyped after they had laid F3 generation progeny. In parallel, we crossed a wild-type male with a wild-type hermaphrodite to generate pure wild-type descendants and control for any beneficial longevity effects that could come from crossing rather than self-mating (Fig. 1a). Longevity of genetically wild-type descendents from wild-type or *wdr-5* mutant ancestors was compared in the F3, F4 and F5 generations. Interestingly, genetically wild-type F3 descendants from P0 *wdr-5* parents (+/+ from P0 *wdr-5* parents) still showed a ~20% extension of lifespan ($P < 0.0001$) compared to descendants from pure wild-type parents (+/+ from P0 WT parents) (Fig. 1b). This 20% lifespan extension was similar in magnitude to the lifespan extension of pure F3 *wdr-5*(*ok1417*) mutants (*wdr-5*/*wdr-5*) (Fig. 1b). The lifespan of genetically wild-type descendants from *wdr-5*(*ok1417*) mutant parents (+/+ from P0 *wdr-5* parents) was still extended in the F4 generation (Fig. 1c), but was no longer extended in the F5 generation (Fig. 1d). Thus, *wdr-5* deficiency only in the parental generation can extend the lifespan of subsequent generations. Because the lifespan of F5 generation wild-type descendants from *wdr-5* mutant parents is no longer extended, the lifespan extension observed in the F3 and F4 generations is unlikely to be due to extraneous mutations that might have been present in the parental *wdr-5* mutant strain. Instead, the transgenerational inheritance of longevity may be due to epigenetic changes in H3K4me3 itself or in another molecule that can only be inherited for a limited number of generations.

We next asked if a transgenerational epigenetic heritability of lifespan was also observed with SET-2, the H3K4me3 methyltransferase enzyme that functions together with ASH-2 and WDR-5 to regulate H3K4me3 levels[12,20–22] and longevity in *C. elegans*[12] (Fig. 2). Similar to what we observed for *wdr-5*, genetically wild-type descendants from *set-2*(*ok952*) mutants still had a ~30% extension of lifespan ($P < 0.0001$) in the F3 and F4 generations (Fig. 2b, c), but not in the F5 generation (Fig. 2d). Genetically wild-type F3 descendants from the reverse cross—P0 *set-2*(*ok952*) males crossed with wild-type hermaphrodites—were also long-lived (Supplementary Table 1), indicating that transgenerational inheritance of longevity is not linked to a particular gender in the parental generation.

ASH-2 is important for the conversion of H3K4 dimethylation (H3K4me2) to H3K4me3 (ref. 23). Knockdown of *ash-2* by RNA interference (RNAi) in worms decreases global H3K4me3 levels at the L3 stage[12,22] and extends longevity[12]. We asked if *ash-2* knockdown only in the parental generation affected the lifespan of several generations of descendants. Wild-type parent worms (P0) were placed on plates with bacteria expressing RNAi to *ash-2* from birth to the larval stage L4, then switched every day for 3 days onto plates

[1]Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA. [2]Cell Biology Department, Harvard Medical School and Division of Newborn Medicine, Children's Hospital, 300 Longwood Avenue, Boston, Massachusetts 02115, USA.

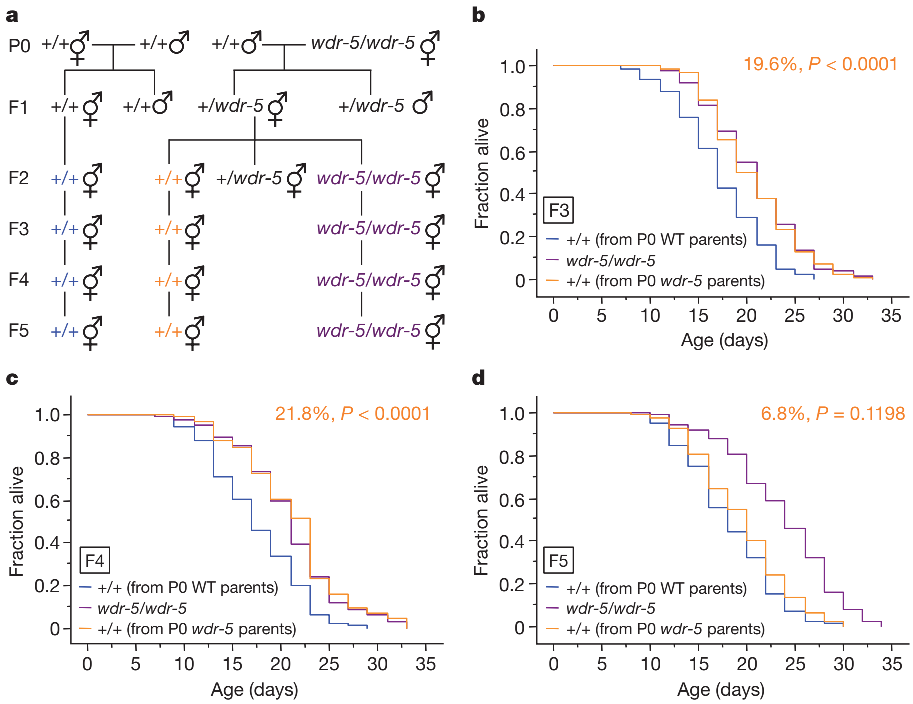


Figure 1 | Genetically wild-type descendants from *wdr-5* mutant parents have extended lifespan for several generations. **a**, Scheme for generating wild-type (+/+) descendants from *wdr-5*(*ok1417*) mutant worms (*wdr-5*/*wdr-5*). **b**–**d**, Lifespan of genetically wild-type F3 (**b**), F4 (**c**) and F5 (**d**) descendants of *wdr-5*(*ok1417*) mutant worms (+/+ from P0 *wdr-5* parents) compared to descendants of wild-type worms (+/+ from P0 WT parents). Mean lifespan and statistics are presented in Supplementary Table 1.

containing OP50-1 bacteria and streptomycin to selectively prevent the growth of RNAi-expressing bacteria (Fig. 3a). Endogenous *ash-2* messenger RNA and ASH-2 protein levels were significantly decreased in the P0 generation, but returned to normal levels in subsequent generations (Fig. 3b, c), indicating that *ash-2* RNAi is not itself inherited. The lifespan of worms from the F1, F2 and F3 generations in which *ash-2* had been knocked down only in the P0 parental generation was still significantly extended (19–27%, $P < 0.0001$) compared to that of descendants of worms treated with empty vector control in the P0 parental generation (Fig. 3d–g). By contrast, F4 generation descendants no longer had extended lifespan (Fig. 3h). We obtained similar results after bleaching P0 worms to avoid potential carry over of RNAi-expressing bacteria (data not shown). Thus, alteration of the components of the H3K4me3

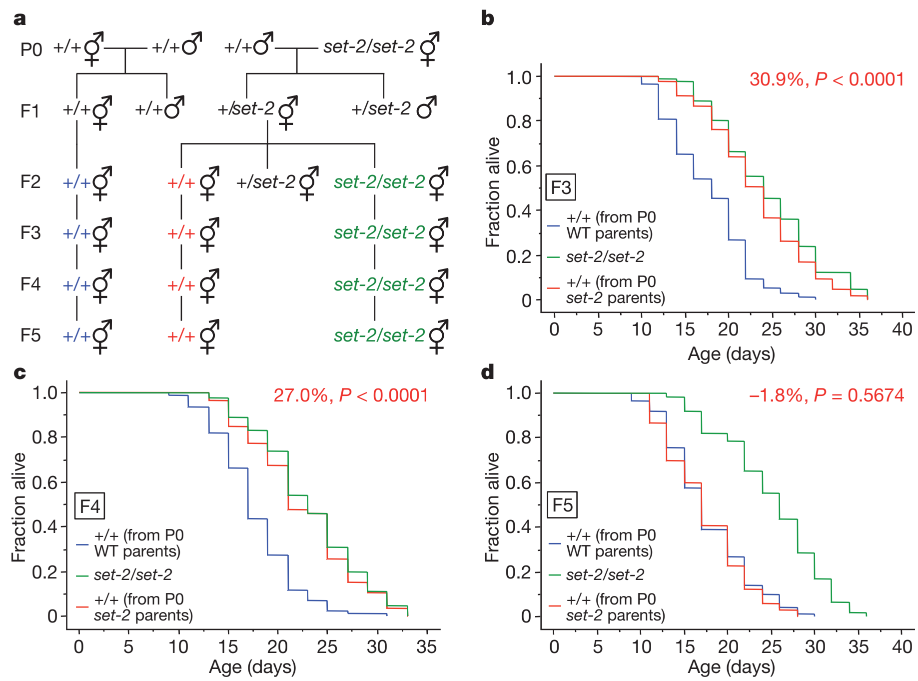


Figure 2 | Genetically wild-type descendants from *set-2* mutant parents have extended lifespan for several generations. **a**, Scheme for generating wild-type (+/+) descendants from *set-2*(*ok952*) mutant worms (*set-2*/*set-2*). **b**–**d**, Lifespan of genetically wild-type F3 (**b**), F4 (**c**) and F5 (**d**) descendants from *set-2*(*ok952*) mutant worms (+/+ from P0 *set-2* parents) compared to descendants of wild-type worms (+/+ from P0 WT parents). Mean lifespan and statistics are presented in Supplementary Table 1.

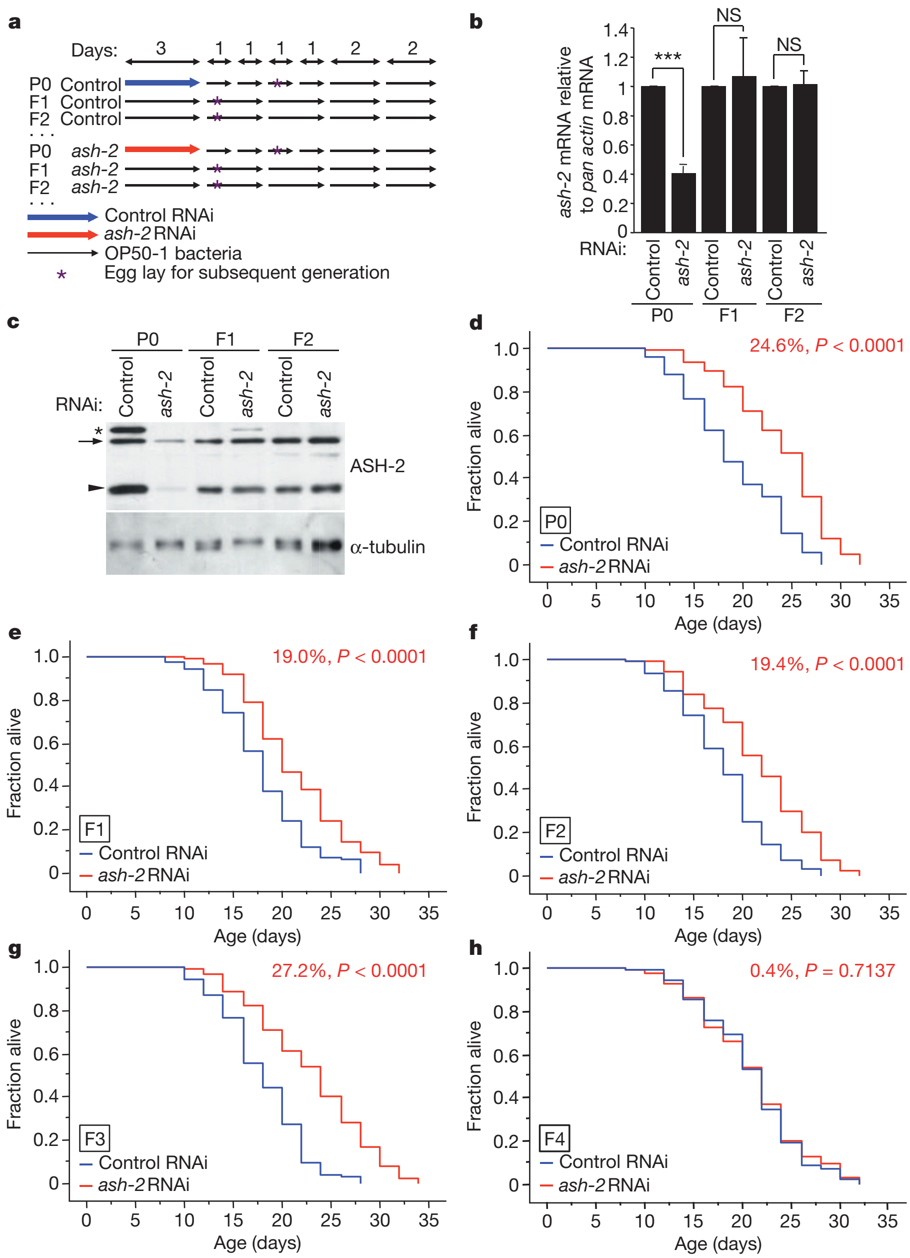


Figure 3 | Knockdown of *ash-2* only in the parental generation extends lifespan for several generations. a, Scheme for generating wild-type descendants from RNAi-treated parents. **b**, *ash-2* mRNA levels at day 7 in different generations of worms treated with *ash-2* RNAi or empty vector (control) only in the P0 generation. Mean ± s.e.m. of three independent experiments. ***$P = 0.0002$ with paired *t*-test. **c**, ASH-2 protein levels at L3 stage in different generations of worms treated with *ash-2* RNAi or empty vector (control) only in the P0 generation. Representative of two independent experiments. *, non-specific band; arrow, ASH-2; arrowhead, protein related to ASH-2, possibly a degradation product. **d**–**h**, Lifespan of P0 (**d**), F1 (**e**), F2 (**f**), F3 (**g**) and F4 (**h**) generations of worms with RNAi knockdown of *ash-2* or control RNAi (empty vector) in parents only. Mean lifespan and statistics are presented in Supplementary Table 2.

methyltransferase complex (ASH-2, WDR-5 and SET-2) in parents affects the lifespan of descendants, supporting the possibility that transgenerational inheritance of longevity is due to epigenetic changes that may only be inherited for a limited number of generations.

Importance of the H3K4me3 demethylase and germline

The H3K4me3 demethylase RBR-2 is necessary for the lifespan extension caused by deficiencies in members of the ASH-2 complex[12]. We asked if the transgenerational extension of longevity induced by deficiencies in members of the ASH-2 complex is dependent on RBR-2. The lifespan of genetically wild-type F3 descendants from P0 *wdr-5* parents (+/+ from P0 *wdr-5* parents) was no longer extended in the presence of *rbr-2* RNAi (Fig. 4a, b). Similarly, F3 wild-type descendants from *set-2;rbr-2* parents (+/+ from P0 *set-2;rbr-2* parents) were no longer long-lived (Supplementary Fig. 1). Together, these data indicate that the transgenerational inheritance of longevity due to deficiencies in H3K4 trimethylation complex members is dependent on the H3K4me3 demethylase RBR-2. The fact that the longevity of wild-type descendants of *wdr-5* and *set-2* mutants is reverted by deficiencies in *rbr-2* also indicates that this extended lifespan is unlikely to result from extraneous mutations in *wdr-5* or *set-2* strains. *rbr-2* mutation or knockdown did not lead to a shortening of lifespan in descendants (Supplementary Fig. 2), indicating that by itself, RBR-2 deficiency does not affect longevity in a transgenerational manner.

Longevity due to modulation of the ASH-2 complex is dependent on a functioning germline[12]. To test if wild-type descendants of worms

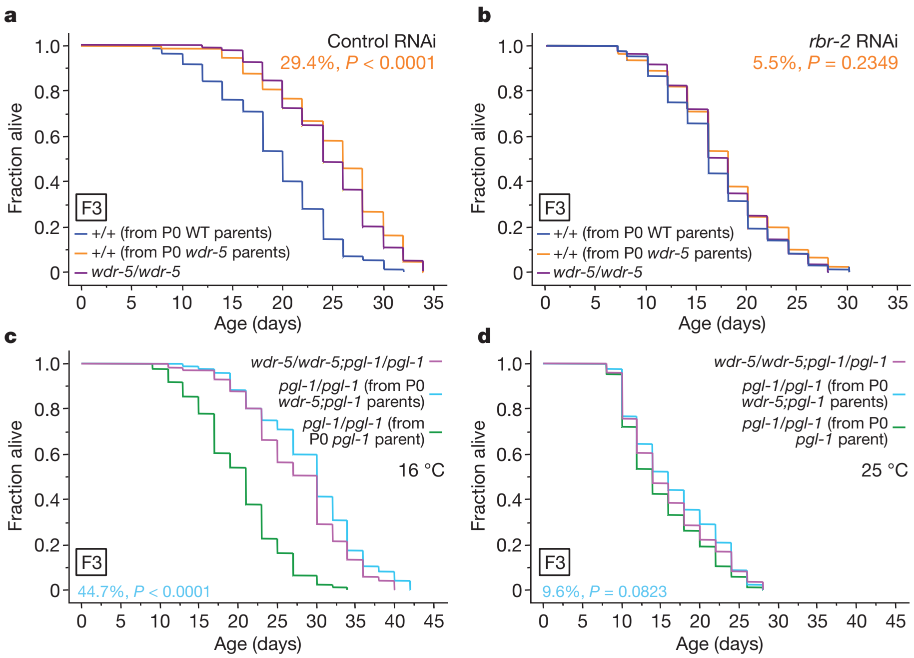


Figure 4 | Transgenerational inheritance of longevity by deficiencies in ASH-2 complex members is dependent on the presence of the H3K4me3 demethylase RBR-2 and an intact germline. **a**, **b**, Lifespan of genetically wild-type F3 descendants from *wdr-5*(*ok1417*) mutant worms (+/+ from P0 *wdr-5* parents) in the presence of empty vector (control RNAi) (**a**) or *rbr-2* RNAi (**b**). **c**, **d**, Lifespan of *pgl-1* F3 descendants from *wdr-5*(*ok1417*);*pgl-1*(*bn101*) mutant worms (*pgl-1/pgl-1* from P0 *wdr-5;pgl*-1 parents) compared with descendants from *pgl-1*(*bn101*) worms at the permissive temperature (16 °C) (**c**) and at the restrictive temperature (25 °C) (**d**). Mean lifespan and statistics are presented in Supplementary Tables 3 and 4.

with deficiencies in ASH-2 complex members also require the presence of a functioning germline for lifespan extension, we used temperature-sensitive feminized *fem-3*(*e2006*) mutant worms, which do not produce mature eggs at the restrictive temperature[24]. Knockdown of *ash-2* and *wdr-5* only in parents extended the lifespan of the F1 generation in *fem-3*(*e2006*) mutant worms at the permissive temperature (16 °C), but not at the restrictive temperature (25 °C) (Supplementary Fig. 3). To independently examine if the germline is required for the longevity of wild-type descendants of mutants of ASH-2 complex members, we used *pgl-1*(*bn101*) temperature-sensitive mutants that cannot form a functioning germline at the restrictive temperature[25] (Fig. 4c, d). F3 generation *pgl-1* descendants from *wdr-5;pgl-1* mutant parents no longer had an extended lifespan compared to *pgl-1* descendants from *pgl-1* parents at the restrictive temperature (25 °C) (Fig. 4c, d). Thus, a functioning adult germline is necessary for the long lifespan of wild-type descendants of parents with deficiencies in members of the ASH-2 complex.

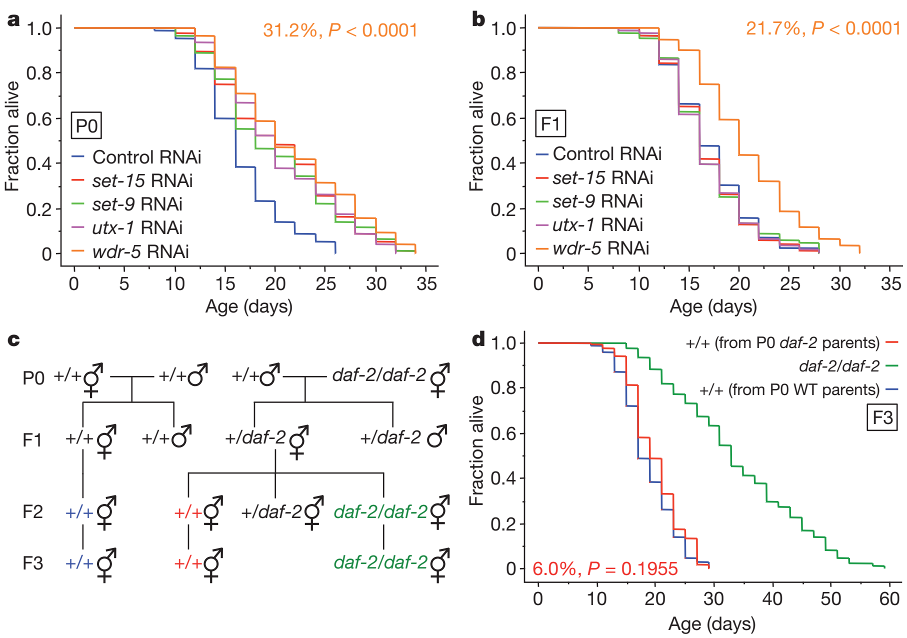


Figure 5 | Other longevity regulators do not have a transgenerational effect on lifespan. **a**, **b**, Lifespan of P0 (**a**) and F1 (**b**) generation descendants from worms treated with *set-9*, *set-15*, *utx-1* and *wdr-5* RNAi or control RNAi (empty vector) only in the P0 generation. **c**, Scheme for generating wild-type (+/+) progeny from *daf-2*(*e1370*) mutant worms (*daf-2/daf-2*). **d**, Lifespan of genetically wild-type F3 descendants from *daf-2*(*e1370*) mutant worms (+/+ from P0 *daf-2* parents). Mean lifespan and statistics are presented in Supplementary Tables 2 and 4.

Specificity of epigenetic memory of lifespan

We then asked if the transgenerational inheritance of longevity is specific to H3K4me3 modifiers or if it is also observed with chromatin modifiers of other marks (*set-9*, *set-15* and *utx-1*), and more generally with genes in known longevity pathways: insulin signalling (*age-1* and *dod-23*), mitochondria (*cco-1* and *cyc-1*) and stress resistance (*asm-3*)[12,17,18,26–32]. In contrast to what we observed for *ash-2* and *wdr-5*, knockdown of *set-9*, *set-15*, *utx-1*, *age-1*, *asm-3*, *cco-1*, *cyc-1* and *dod-23* only in parents did not extend the lifespan of the F1 generation (Fig. 5a, b and Supplementary Fig. 5). Similarly, genetically wild-type F3 descendants from long-lived *daf-2*(*e1370*)[33] mutant worms (+/+ from P0 *daf-2* parents) had no significant extension of lifespan (6% $P = 0.1955$) (Fig. 5c, d). Collectively, these findings indicate that transgenerational extension of longevity is relatively specific to H3K4me3 chromatin modifiers, and further indicate that the H3K4me3 mark may be important for epigenetic memory of lifespan between generations. As SET-9, SET-15 and UTX-1, unlike members of the ASH-2 complex, regulate lifespan in a manner that is independent of the germline[12,17,18], it is also possible that transgenerational inheritance of longevity is specific to chromatin regulators that act in the germline.

Transgenerational inheritance of gene expression

We next determined if transgenerational inheritance of lifespan is associated with heritable changes in H3K4me3. Western blot and immunocytochemistry showed that global H3K4me3 levels were not decreased in F3 and F4 generation genetically wild-type descendants from *wdr-5* and *set-2* parents or in F1 and F2 generation descendants from *ash-2* or *wdr-5* knockdown only in parents (Fig. 6a, b and Supplementary Figs 6, 7). Thus, transgenerational inheritance of lifespan is unlikely to be mediated by a heritable global decrease in H3K4me3 levels. Transgenerational inheritance of lifespan might be associated with heritable local changes of H3K4me3 at certain loci, which could affect expression of certain genes involved in longevity. To test this idea, we compared gene expression genome-wide in wild-type descendants from *wdr-5* mutant and wild-type ancestors, and

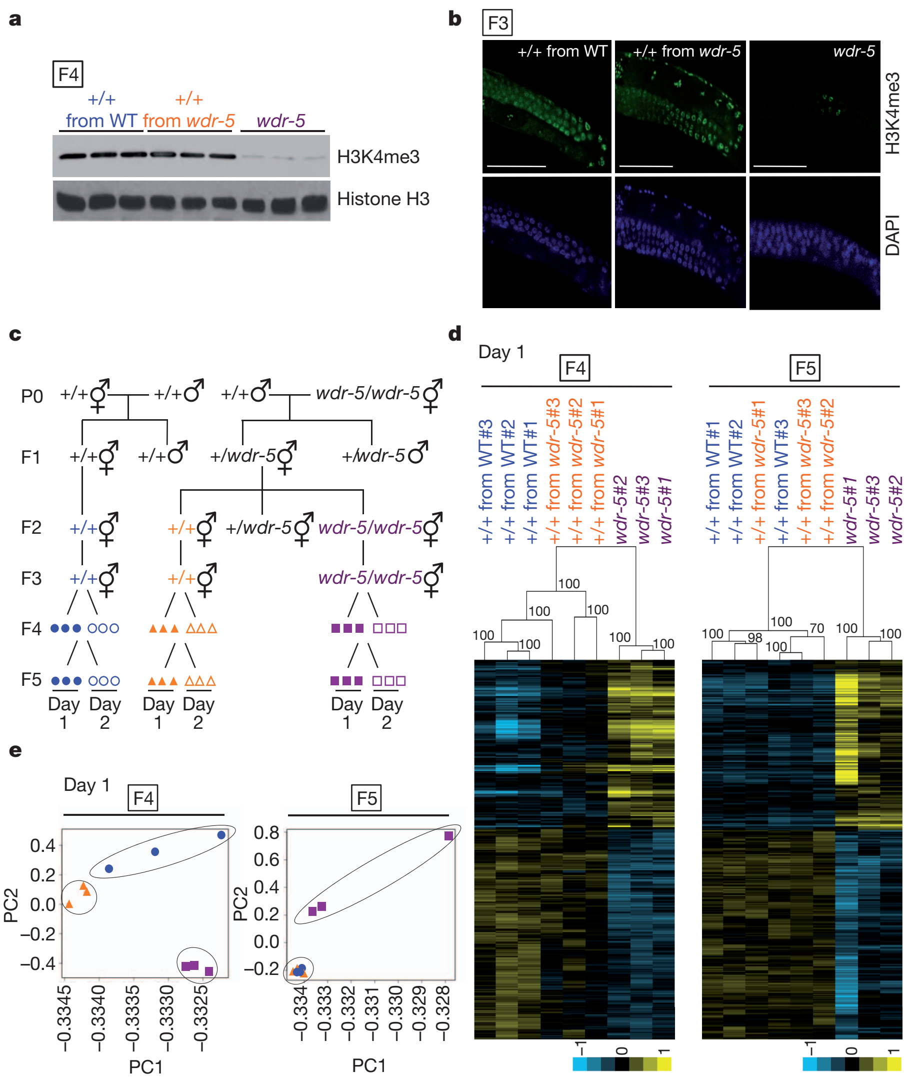


Figure 6 | Genetically wild-type descendants from *wdr-5* mutant parents exhibit differences in gene expression, but not in global H3K4me3 levels, compared to descendants from wild-type parents. **a**, **b**, Global H3K4me3 levels in the F4 generation by western blot (**a**) or in the F3 generation by immunocytochemistry (**b**) of L3 worms from genetically wild-type descendants from *wdr-5* parents (+/+ from *wdr-5*) or wild-type parents (+/+ from WT), and *wdr-5* mutants (*wdr-5*). Scale bars, 50 μm. **c**, Scheme for generating wild-type descendants from a cross between *wdr-5*(*ok1417*) null mutant worms and wild-type worms. Symbols represent RNA samples from L3 worms from three independent F2 ancestors on the first (closed symbols) or second (open symbols) day of egg-laying. **d**, Unbiased hierarchical clustering of WDR-5 regulated genes from the first day of egg-laying (Supplementary Table 9). Pvclust values are displayed on each node of the dendrogram. Values superior to 95 are considered significant. **e**, Principal component analysis (PCA) of the entire microarray data sets from the first day of egg-laying (Supplementary Table 5). PC, principal component. Symbols represent gene expression data from L3 worms collected on the first day of egg-laying (Fig. 6c).

pure *wdr-5* mutant descendants in the F4 and F5 generations (Fig. 6c). For each condition, we collected triplicates of L3 stage worms from the first or second day of egg-laying (Fig. 6c), with the first day of egg-laying corresponding to the samples used for lifespan assays. Statistical analysis of microarray (SAM) identified 759 genes that were differentially regulated in *wdr-5* pure mutants compared to wild-type worms, regardless of the generation (Supplementary Table 7) and egg-laying day (Supplementary Fig. 8a) ($P = 2.38 \times 10^{-116}$, hypergeometric probability). These WDR-5-regulated genes are enriched for longevity, development and growth gene ontology terms (Supplementary Fig. 8b), consistent with WDR-5's reported functions[12,21,22]. As expected, WDR-5-regulated genes significantly overlap with ASH-2-regulated genes[12] ($P = 6.14 \times 10^{-12}$, hypergeometric probability, Supplementary Fig. 8c) and are enriched for H3K4me3 (refs 34, 35) ($P = 2.49 \times 10^{-34}$, hypergeometric probability, Supplementary Fig. 8d). These observations indicate that WDR-5 functions together with ASH-2 to regulate a subset of genes by modulating H3K4me3 at these loci.

We asked if the expression of some WDR-5-regulated genes might be transgenerationally inherited. Interestingly, a significant subset of WDR-5-regulated genes was still differentially regulated in wild-type descendants from *wdr-5* mutant worms in the F4 generation (Fig. 6d and Supplementary Fig. 10a), but not in the F5 generation (Fig. 6d and Supplementary Fig. 10a), consistent with the return to a normal lifespan in the F5 generation. Unbiased hierarchical clustering analysis showed that WDR-5-regulated genes in wild-type descendants from *wdr-5* mutant versus wild-type parents still clustered separately in the F4, but not the F5 generation (Fig. 6d and Supplementary Fig. 10a). Principal component analysis (PCA) confirmed that overall gene expression in wild-type descendants from *wdr-5* parents versus wild-type parents is easily distinguishable in the F4, but not the F5 generation (Fig. 6e and Supplementary Fig. 10b). Genes with transgenerational inheritance of expression were slightly more enriched for H3K4me3 than expected by chance ($P = 0.0123$ and $P = 0.0769$ for the first and second day of egg-laying, respectively, hypergeometric probability), and may represent the genes that are the most affected by the loss of the H3K4me3 mark. A number of these genes are known longevity regulators and are expressed in the germline (Supplementary Table 7). Gene ontology analysis of genes with transgenerational inheritance of expression shows enrichment for different types of metabolic pathways (Supplementary Figs 9, 10c), raising the possibility that changes in metabolism may play a role in the heritability of the phenotype. The genes with transgenerational inheritance of expression were different on the first versus second day of egg-laying, and were no longer identified when samples from different days of egg-laying were pooled (E.L.G. and A.B., data not shown). This could be because worms produced on the first day of egg-laying might be more susceptible to H3K4me3 depletion, because each collection day may represent a different snapshot in the rapidly changing L3 stage[36], or because of inherent stochasticity in the transgenerational inheritance of gene expression. Overall, these results indicate that ancestral H3K4me3 status influences the gene expression of descendants for several generations.

Discussion

Our study provides the first example of epigenetic inheritance of longevity. Histone methylation marks and DNA methylation are generally, but not always, erased between generations with epigenetic reprogramming[37,38]. Our observations are consistent with the notion that H3K4me3 at specific loci may not be completely erased and replenished. Alternatively, the ASH-2/WDR-5/SET-2 complex could control the expression of the genes responsible for the erasure and replenishment of histone methylation marks between generations. Modulation of H3K4me3 modifiers in parents may also affect an unidentified protein or RNA that could in turn be inherited and cause lifespan changes. Interestingly, H3K4me3 regulators have been suggested to have a role in the inheritance of eye colour in *Drosophila*[5,6] and of active transcriptional states in *Dictyostelium*[39]. As the ASH-2 H3K4me3 regulatory complex is conserved from yeast to humans, manipulations of this complex in parents might have a heritable effect on longevity in mammals.

METHODS SUMMARY

Genetic crosses. For genetic crosses, wild-type or *pgl-1* male siblings were crossed to hermaphrodite siblings (to generate wild-type descendants) or to hermaphrodite mutants (to generate wild-type and mutant descendants). F1 hermaphrodites were allowed to lay progeny and were then genotyped by single-worm genotyping to ensure that they were heterozygous at the loci of the mutations of interest. Twenty F2 worms were placed on individual plates and allowed to lay ~40 eggs. F2 parents were then genotyped by single-worm genotyping and the lifespan of three independent lines was analysed. Progeny from each of these independent lines were collected after synchronized egg-laying for the F4 and F5 generations and analysed in lifespan assays. For microarray analysis, RNA samples were isolated at the F4 and F5 generations at the larval stage L3. Each sample was prepared from independent F2 clonal parents in triplicate with ~1,000 worms each.

Lifespan assays. Worm lifespan assays were performed at 20 °C, unless noted differently, without 5-fluoro-2′-deoxyuridine (FUdR), as described previously[40]. For each lifespan assay, 90 worms per condition were used in three plates (30 worms per plate), unless noted differently. Worms that underwent 'matricide', that exhibited ruptured vulva, or crawled off the plates were censored. Statistical analyses of lifespan were performed on Kaplan–Meier survival curves in StatView 5.0.01 by log rank (Mantel–Cox) tests. The values from the Kaplan–Meier curves are included in the Supplementary Tables.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 September 2010; accepted 26 September 2011.
Published online 19 October 2011.

1. Cubas, P., Vincent, C. & Coen, E. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature* **401,** 157–161 (1999).
2. Brink, R. A. A genetic change associated with the *R* locus in maize which is directed and potentially reversible. *Genetics* **41,** 872–889 (1956).
3. Woodhouse, M. R., Freeling, M. & Lisch, D. Initiation, establishment, and maintenance of heritable MuDR transposon silencing in maize are mediated by distinct factors. *PLoS Biol.* **4,** e339 (2006).
4. Katz, D. J., Edwards, T. M., Reinke, V. & Kelly, W. G. A *C. elegans* LSD1 demethylase contributes to germline immortality by reprogramming epigenetic memory. *Cell* **137,** 308–320 (2009).
5. Cavalli, G. & Paro, R. The *Drosophila Fab-7* chromosomal element conveys epigenetic inheritance during mitosis and meiosis. *Cell* **93,** 505–518 (1998).
6. Cavalli, G. & Paro, R. Epigenetic inheritance of active chromatin after removal of the main transactivator. *Science* **286,** 955–958 (1999).
7. Seong, K. H., Li, D., Shimizu, H., Nakamura, R. & Ishii, S. Inheritance of stress-induced, ATF-2-dependent epigenetic change. *Cell* **145,** 1049–1061 (2011).
8. Morgan, H. D., Sutherland, H. G., Martin, D. I. & Whitelaw, E. Epigenetic inheritance at the agouti locus in the mouse. *Nature Genet.* **23,** 314–318 (1999).
9. Blewitt, M. E., Vickaryous, N. K., Paldi, A., Koseki, H. & Whitelaw, E. Dynamic reprogramming of DNA methylation at an epigenetically sensitive allele in mice. *PLoS Genet.* **2,** e49 (2006).
10. Rassoulzadegan, M. *et al.* RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature* **441,** 469–474 (2006).
11. Dang, W. *et al.* Histone H4 lysine 16 acetylation regulates cellular lifespan. *Nature* **459,** 802–807 (2009).
12. Greer, E. L. *et al.* Members of the H3K4 trimethylation complex regulate lifespan in a germline-dependent manner in *C. elegans*. *Nature* **466,** 383–387 (2010).
13. Siebold, A. P. *et al.* Polycomb Repressive Complex 2 and Trithorax modulate *Drosophila* longevity and stress resistance. *Proc. Natl Acad. Sci. USA* **107,** 169–174 (2010).
14. McColl, G. *et al.* Pharmacogenetic analysis of lithium-induced delayed aging in *Caenorhabditis elegans*. *J. Biol. Chem.* **283,** 350–357 (2008).
15. Chen, S. *et al.* The conserved NAD(H)-dependent corepressor CTBP-1 regulates *Caenorhabditis elegans* life span. *Proc. Natl Acad. Sci. USA* **106,** 1496–1501 (2009).
16. Takahashi, Y. *et al.* Asymmetric arginine dimethylation determines life span in *C. elegans* by regulating forkhead transcription factor DAF-16. *Cell Metab.* **13,** 505–516 (2011).
17. Maures, T. J., Greer, E. L., Hauswirth, A. G. & Brunet, A. The H3K27 demethylase UTX-1 regulates *C. elegans* lifespan in a germline-independent, insulin-dependent manner. *Aging Cell* doi: 10.1111/j.1474-9726.2011.00738.x (11 August 2011).
18. Jin, C. *et al.* Histone demethylase UTX-1 regulates *C. elegans* life span by targeting the insulin/IGF-1 signaling pathway. *Cell Metab.* **14,** 161–172 (2011).
19. Steward, M. M. *et al.* Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nature Struct. Mol. Biol.* **13,** 852–854 (2006).
20. Simonet, T., Dulermo, R., Schott, S. & Palladino, F. Antagonistic functions of SET-2/SET1 and HPL/HP1 proteins in *C. elegans* development. *Dev. Biol.* **312,** 367–383 (2007).

21. Xiao, Y. *et al. Caenorhabditis elegans* chromatin-associated proteins SET-2 and ASH-2 are differentially required for histone H3 Lys 4 methylation in embryos and adult germ cells. *Proc. Natl Acad. Sci. USA* **108,** 8305–8310 (2011).
22. Li, T. & Kelly, W. G. A role for Set1/MLL-related components in epigenetic regulation of the *Caenorhabditis elegans* germ line. *PLoS Genet.* **7,** e1001349 (2011).
23. Dou, Y. *et al.* Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nature Struct. Mol. Biol.* **13,** 713–719 (2006).
24. Haag, E. S., Wang, S. & Kimble, J. Rapid coevolution of the nematode sex-determining genes *fem-3* and *tra-2*. *Curr. Biol.* **12,** 2035–2041 (2002).
25. Kawasaki, I. *et al.* PGL-1, a predicted RNA-binding component of germ granules, is essential for fertility in *C. elegans*. *Cell* **94,** 635–645 (1998).
26. Fisher, K., Southall, S. M., Wilson, J. R. & Poulin, G. B. Methylation and demethylation activities of a *C. elegans* MLL-like complex attenuate RAS signalling. *Dev. Biol.* **341,** 142–153 (2010).
27. Curran, S. P. & Ruvkun, G. Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS Genet.* **3,** e56 (2007).
28. Hamilton, B. *et al.* A systematic RNAi screen for longevity genes in *C. elegans*. *Genes Dev.* **19,** 1544–1555 (2005).
29. Kim, Y. & Sun, H. Functional genomic approach to identify novel genes involved in the regulation of oxidative stress resistance and animal lifespan. *Aging Cell* **6,** 489–503 (2007).
30. Lee, S. S. *et al.* A systematic RNAi screen identifies a critical role for mitochondria in *C. elegans* longevity. *Nature Genet.* **33,** 40–48 (2003).
31. Dillin, A. *et al.* Rates of behavior and aging specified by mitochondrial function during development. *Science* **298,** 2398–2401 (2002).
32. Murphy, C. T. *et al.* Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* **424,** 277–283 (2003).
33. Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A *C. elegans* mutant that lives twice as long as wild type. *Nature* **366,** 461–464 (1993).
34. Liu, T. *et al.* Broad chromosomal domains of histone modification patterns in *C. elegans. Genome Res.* **21,** 227–236 (2011).
35. Gerstein, M. B. *et al.* Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330,** 1775–1787 (2010).
36. Spencer, W. C. *et al.* A spatial and temporal map of *C. elegans* gene expression. *Genome Res.* **21,** 325–341 (2011).
37. Wu, S. C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nature Rev. Mol. Cell Biol.* **11,** 607–620 (2010).
38. Martin, C. & Zhang, Y. Mechanisms of epigenetic inheritance. *Curr. Opin. Cell Biol.* **19,** 266–272 (2007).
39. Muramoto, T., Muller, I., Thomas, G., Melvin, A. & Chubb, J. R. Methylation of H3K4 Is required for inheritance of active transcriptional states. *Curr. Biol.* **20,** 397–406 (2010).
40. Greer, E. L. *et al.* An AMPK-FOXO pathway mediates longevity induced by a novel method of dietary restriction in *C. elegans*. *Curr. Biol.* **17,** 1646–1656 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to J. Lieb, A. Rechtsteiner and S. Strome for sharing their ModENCODE data pre-publication and for helpful discussion. We thank K. Shen, M. W. Tan and T. Stiernagle and the *Caenorhabditis* Genetics Center for gifts of strains and reagents. We thank B. Meyer for her gift of the ASH-2 antibody. We thank A. Fire, S. Kim, J. Sage, S. Iwase, J. Lipsick, E. Pollina, A. Villeneuve and members of the Brunet lab for discussions and critical reading of the manuscript. We thank S. Han for screening different H3K4me3 antibodies for western blots in worm extracts. We thank R. Liefke and H. Tang for help with microarray analysis. This work was supported by NIH R01-AG31198 grant and by a generous gift from the Glenn Foundation for Medical Research to A.B.; E.L.G. was supported by an NSF graduate fellowship, by NIH ARRA-AG31198, by T32-CA009361, by a Helen Hay Whitney Post-Doctoral fellowship, and by a NIH R01-GM058012 (to Y.S.). T.J.M. was supported by NIH F32-AG037254. J.P.L. was supported by NIH T32-MH020016.

Author Contributions E.L.G. conceived and planned the study with the help of A.B. E.L.G. performed the experiments and wrote the paper with the help of A.B.; E.L.G. performed some of the experiments in the lab of Y.S.; T.J.M. performed immunocytochemistry experiments (Fig. 6b and Supplementary Figs 6c and 7c); D.U. performed Pvclust and PCA microarray analysis (Fig. 6d, e and Supplementary Fig. 10a, b). A.G.H. helped with Figs 3b, c and 6a and Supplementary Figs 6a, b and 7a, b. E.M. performed an independent repeat of the transgenerational *wdr-5* RNAi longevity experiments (Supplementary Table 2). J.P.L. helped with Fig. 3c and Supplementary Fig. 7a, b. B.A.B. helped with bioinformatics analysis (Supplementary Table 7). All authors discussed the results and commented on the manuscript.

Author Information The raw unfiltered microarray results are deposited at the Gene Expression Omnibus (GEO) under the Subseries entry GSE31043. The raw unfiltered chromatin immunoprecipitation (ChIP)-chip data are deposited at GEO under the Subseries entry GSE30789. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.B. (anne.brunet@stanford.edu).

METHODS

Worm strains. *wdr-5*(*ok1417*) and *set-2*(*ok952*) strains were provided by the *Caenorhabditis* Genetics Center. Wild-type (N2), *wdr-5*(*ok1417*) and *set-2*(*ok952*) strains were genotyped for *mut-16*(*mg461*), a mutation that affects RNAi efficiency and that was found as an extraneous mutation in several laboratory strains. These strains did not contain the *mut-16*(*mg461*) mutation. *wdr-5*(*ok1417*) mutant worms were backcrossed four to nine times by crossing wild-type males with *wdr-5*(*ok1417*) hermaphrodites. *set-2*(*ok952*) mutant worms were backcrossed four to six times by crossing wild-type males with *set-2*(*ok952*) hermaphrodites. The transgenerational inheritance of longevity was similar, both in terms of magnitude and number of generation, whether *wdr-5*(*ok1417*) and *set-2*(*ok952*) worms were backcrossed four to nine times or four to six times, respectively (see Supplementary Table 1), arguing against a simple backcrossing effect to explain the increased lifespan of wild-type descendants from *wdr-5* or *set-2* parents. *rbr-2*(*tm1231*) mutant worms were backcrossed seven times, *daf-2*(*e1370*) mutant worms were backcrossed an additional two times by our lab, *set-2*(*ok952*) and *rbr-2*(*tm1231*) were backcrossed two times and six times, respectively, before being crossed together to generate *set-2*(*ok952*);*rbr-2*(*tm1231*) and then crossed an additional time to wild-type worms; *wdr-5*(*ok1417*) and *pgl-1*(*bn101ts*) were backcrossed four times each before being crossed together to generate *wdr-5*(*ok1417*);*pgl-1*(*bn101ts*) and then crossed an additional time to *pgl-1*(*bn101ts*) worms. For crosses involving *set-2*(*ok952*);*rbr-2*(*tm1231*) mutant worms, six F3 progeny were genotyped for each independent line to ensure the genotype of the second mutant loci. Temperature-sensitive *fem-3*(*e2006*) mutant worms were either maintained at 16 °C or were switched to 25 °C at birth and maintained at this temperature for the entirety of their lifespan. Temperature-sensitive *pgl-1*(*bn101*) and *wdr-5*(*ok1417*);*pgl-1*(*bn101*) mutant worms were either maintained at 16 °C or were switched to 25 °C at the L4 stage in F2 parents. F3 progeny from these worms was maintained at 16 °C or 25 °C for the entirety of their lifespan.

RNA interference. Adult worms were placed on NGM plates containing ampicillin (100 mg ml^{-1}) and IPTG (0.4 mM) seeded with the respective bacteria and removed after 4–6 h to obtain synchronized populations of worms. HT115 (DE3) bacteria transformed with vectors expressing RNAi to the genes of interest were all obtained from the Ahringer library (a gift from M. W. Tan), except RNAi to *rbr-2* that was from the Open Biosystems library (a gift from K. Shen). At the L4 stage, P0 worms were moved to NGM plates containing streptomycin (300 µg ml^{-1}) seeded with OP50-1 bacteria, which are streptomycin-resistant, to eliminate any potentially remaining RNAi HT115 (DE3) bacteria, which are streptomycin-sensitive. P0 worms were switched to fresh OP50-1 seeded plates every day until day 6 of life (day 2 of adulthood). Day 6 P0 worms were allowed to lay eggs for 4–6 h and progeny from that stage were picked from these plates and their lifespan was examined. Subsequent generations were obtained by placing young adult F1, F2, F3, or F4 worms on fresh OP50-1 seeded plates for 4–6 h. To perform RNAi in *fem-3*(*e2006*) mutant worms, one set of P0 worms was maintained at 16 °C to allow them to lay eggs for the F1 generation, while a second set of P0 worms was analysed in lifespan assays at both 16 °C and 25 °C. RNAi to *rbr-2* was initiated at the eggs or L1 stage of F3 generation *wdr-5*(*ok1417*) mutant worms, wild-type descendants of *wdr-5*(*ok1417*) mutant worms, and wild-type descendants of wild-type parents.

Real-time quantitative reverse transcription polymerase chain reaction (qRT–PCR). Two hundred worms were picked to NGM plates with OP50-1 bacteria overnight 2 days in a row. Worms were then picked to NGM plates without bacteria and washed three times with M9 buffer (22 mM KH_2PO_4, 34 mM K_2HPO_4, 86 mM NaCl, 1 mM $MgSO_4$). Worm pellets were resuspended in TRIzol (Invitrogen), followed by six freeze-thaw cycles in liquid N_2. One µg of total RNA was reverse-transcribed with oligo dT primers using Superscript II reverse transcriptase (Invitrogen) according to the manufacturer's protocol. Real time PCR was performed on a Bio-Rad iCycler or Roche LightCycler 480II using iQ SYBR green (Bio-Rad) or LightCycler480 SYBR green I master (Roche) with the following primers: pan-actin F: 5′-TCGGTATGGGACAGAAGGAC-3′, pan-actin R: 5′-CATCCCAGTTGGTGACGATA-3′, *ash-2* F: 5′-CGATCGAAACACGGAACGA-3′, *ash-2* R: 5′-TGCCGGAATCTGCAGTTTTT-3′, *wdr-5* F: 5′-CCCTGAAACAATACACTGGACACG-3′, *wdr-5* R: 5′-AACTGGATGACAATCGGAGGC-3′. The experiments were conducted in duplicate and the results were expressed as $2^{(-(\text{target gene number of cycles} - \text{pan-actin number of cycles}))}$.

Protein analysis by western blot. Worms were synchronously grown to the L3 stage and washed off plates with M9 buffer (22 mM KH_2PO_4, 34 mM K_2HPO_4, 86 mM NaCl, 1 mM $MgSO_4$). Worms were washed several times in M9 buffer and snap-frozen in liquid N_2. Sample buffer (2.36% SDS, 9.43% glycerol, 5% β-mercaptoethanol, 0.0945 M Tris HCl pH 6.8, 0.001% bromophenol blue) was added to worm pellets and they were repeatedly snap-frozen in liquid N_2. Worm extracts were sonicated three times for 30 s at ~15 W (VirSonic 600) and boiled for 2 min before being resolved on SDS–PAGE (10% or 14%) and transferred to nitrocellulose membranes. The membranes were incubated with primary antibodies (H3K4me3 (Abcam ab8580, Millipore 07-473), 1:500; H3 (Abcam ab1791), 1:1,000; ASH-2 antibody[41] (a gift from B. J. Meyer), 1:2,000, alpha-tubulin (Sigma T9026), 1:1,000), and the primary antibodies were visualized using horseradish-peroxidase-conjugated anti-rabbit secondary antibody (Calbiochem 401393) and ECL Plus (Amersham Biosciences).

Whole-mount immunocytochemistry. Worms were washed several times to remove bacteria and resuspended in fixing solution (160 mM KCl, 100 mM Tris-HCl pH 7.4, 40 mM NaCl, 20 mM Na_2EGTA, 1 mM EDTA, 10 mM spermidine HCl, 30 mM PIPES pH 7.4, 1% Triton X-100, 50% methanol, 2% formaldehyde) and subjected to two rounds of snap freezing in liquid N_2. The worms were fixed at 4 °C for 30 min and washed briefly in T buffer (100 mM Tris HCl pH 7.4, 1 mM EDTA, 1% Triton X-100) before a 1 h incubation in T buffer supplemented with 1% β-mercaptoethanol at 37 °C. The worms were washed with borate buffer (25 mM H_3BO_3, 12.5 mM NaOH pH 9.5) and then incubated in borate buffer containing 10 mM DTT for 15 min. Worms were blocked in PBST (PBS pH 7.4, 0.5% Triton X-100, 1 mM EDTA) containing 1% BSA for 30 min and incubated overnight with H3K4me3 antibody (Millipore 07-473; 1:100) and with Alexa Fluor 594 secondary antibody (Invitrogen; 1:25–1:100). DAPI (2 mg ml^{-1}) was added to visualize nuclei. The worms were mounted on a microscope slide and visualized using a Leica SP2 confocal system or a Zeiss Axioskop2 plus fluorescence microscope.

Single-worm genotyping. Single worms were placed in 5 µl of worm lysis buffer (50 mM KCl, 10 mM Tris pH 8.3, 2.5 mM $MgCl_2$, 0.45% NP40, 0.45% Tween-20, 0.01% gelatin (w/v) and 60 mg ml^{-1} proteinase K), and incubated at -80 °C for 1 h, 60 °C for 1 h, and then 95 °C for 15 min. PCR reactions were performed using the following primers: *set-2* F: 5′-TGAAAGGATGATACTCGTGGGC-3′, *set-2* R: 5′-CGATGAGAGAAAGGGGATTTTGTAAC-3′, *wdr-5* F: 5′-TTGTGTGTTCGCTGTGCATG-3′, *wdr-5* R: 5′-GTATTTGCTCTCGGTCGATC-3′, *mut-16* F: 5′-AATATTCGATCGGCAAGCAG-3′, *mut-16* R: 5′-CCCGCCGATACAGAAACTAA-3′, *rbr-2* F: 5′-CAAGTGTCGTGTGATGCTGTGG-3′, *rbr-2* R: 5′-TGGCGATTGGAAACTCCGAG-3′, *pgl-1* F: 5′-TGATGTGATTGCCGAGGAACAC-3′, *pgl-1* R; 5′-GCTGAAGAAGACTGAAGACGCTAAG-3′, *daf-2* F: 5′-ACCTGGAGTCGCTCAAGTTTTG-3′, *daf-2* R: 5′-TGCTTCGCTTTCATCGGTGTC-3′.

PCR reactions were performed according to the manufacturer's protocol (Qiagen) and PCR reactions were resolved on agarose gels. *daf-2* PCR products were digested with BlpI to distinguish between wild-type and *daf-2*(*e1370*) genotypes. *pgl-1* PCR products were digested with MseI to distinguish between wild-type and *pgl-1*(*bn101*) genotypes.

Microarray analysis. Total RNA was isolated using an RNAqueous kit (Ambion). Microarray hybridization was performed at the Stanford Protein and Nucleic Acid facility with oligonucleotide arrays (Affymetrix, GeneChip *C. elegans* Genome Arrays). The raw unfiltered microarray results are deposited at the Gene Expression Omnibus (GEO) under the Subseries entry GSE31043. Background adjustment and normalization was performed with RMA (Robust Multiarray Analysis). Two-class unpaired analysis in significance analysis of microarrays (SAM)[42] was performed with 100 permutations, a 10^6 seed for the random number generator and a 5% false discovery rate (FDR) to compare gene expression in *wdr-5* mutants and wild-type descendants from wild-type parents. To obtain a 5% FDR, a 1.06 delta value was used for samples collected at the first day of egg-laying (day 1) and a 0.93 delta value for samples collected at the second day of egg-laying (day 2). Significantly changed probes from these two lists were then used to compare wild-type descendants from wild-type parents to wild-type descendants from *wdr-5* parents in each generation using a 5% FDR. To obtain a 5% FDR, a 0.66 delta value was used for the F4 generation at day 1, a 0.09 delta value was used for the F5 generation at day 1, a 0.92 delta value was used for the F4 generation at day 2, and a 0.61 delta value was used for the F5 generation at day 2. Similar results for wild-type descendants from *wdr-5* parents compared to wild-type descendants from wild-type parents were observed when SAM was performed with the entire normalized lists of genes.

Hierarchical clustering. A complete linkage hierarchical clustering on the subset of WDR-5 regulated genes for each day (Supplementary Tables 9 and 10) was performed using Gene Cluster 3.0. Clustering results were analysed further with Java TreeView. Further statistical analysis was performed using Pvclust[43]. For the clustering analysis, genes and then arrays were centred using the mean. The R package Pvclust was used to apply complete linkage hierarchical clustering. As the data were centred, the uncentred Pearson correlation coefficient was used as a similarity measure, which was subsequently modified to dissimilarity by subtracting from 1. Experiments were conducted with 1,000 bootstrap replications.

Principal component analysis. Principal component analysis (PCA)[44] was conducted on the entire normalized lists of genes (Supplementary Tables 5 and 6). The data were scaled to obtain unit variance before conducting the PCA analysis. The Prcomp function in the R package 'Stats' was used. The first and the second principal components (PC1 and PC2) were plotted.

H3K4me3 chromatin immunoprecipitation (ChIP)-chip data set from modENCODE and comparison between data sets. The H3K4me3 ChIP-chip data set was generated by the modENCODE consortium from worms at the L3 stage[34,35]. The data, protocols, and antibody information can be accessed at the modENCODE Data Coordination Center (http://intermine.modencode.org), accession ID 3550. Use of this data set during the publication moratorium period was approved (S. Strome, personal communication). The raw unfiltered ChIP-chip data are deposited at GEO under the Subseries entry GSE30789. H3K4me3 ChIP intensity signals were divided by Input signals, log transformed, centred to mean zero, and scaled to standard deviation one. H3K4me3 enrichment peaks (4493) were called using the program ChIPOTle (ref. 45, http://sourceforge.net/projects/chipotle-2/) with a P-value cut-off of 10^{-20}, window size 500 bp, step size 100 bp, and the Bonferroni P-value correction. A list of gene coordinates (transcript start-end) was obtained from WormBase WS170 (http://www.wormbase.org/). Peaks were mapped to 5,062 genes by identifying the genes that had peaks overlap with their 5′ region (500 bp upstream and downstream from the transcript start site). For comparisons between different data sets, hypergeometric probabilities were calculated using http://stattrek.com/Tables/Hypergeometric.aspx.

41. Pferdehirt, R. R., Kruesi, W. S. & Meyer, B. J. An MLL/COMPASS subunit functions in the *C. elegans* dosage compensation complex to target X chromosomes for transcriptional regulation of gene expression. *Genes Dev.* **25,** 499–515 (2011).
42. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98,** 5116–5121 (2001).
43. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22,** 1540–1542 (2006).
44. Pearson, K. On lines and planes of closest fit to systems of points in space. *Phil. Mag.* **6,** 559–572 (1901).
45. Buck, M. J., Nobel, A. B. & Lieb, J. D. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol.* **6,** R97 (2005).

doi:10.1038/nature10529

Two populations of X-ray pulsars produced by two types of supernova

Christian Knigge[1], Malcolm J. Coe[1] & Philipp Podsiadlowski[2]

Two types of supernova are thought to produce the overwhelming majority of neutron stars in the Universe[1]. The first type, iron-core-collapse supernovae, occurs when a high-mass star develops a degenerate iron core that exceeds the Chandrasekhar limit[2]. The second type, electron-capture supernovae, is associated with the collapse of a lower-mass oxygen–neon–magnesium core as it loses pressure support owing to the sudden capture of electrons by neon and/or magnesium nuclei[3,4]. It has hitherto been impossible to identify the two distinct families of neutron stars produced in these formation channels. Here we report that a large, well-known class of neutron-star-hosting X-ray pulsars is actually composed of two distinct subpopulations with different characteristic spin periods, orbital periods and orbital eccentricities. This class, the Be/X-ray binaries, contains neutron stars that accrete material from a more massive companion star[5]. The two subpopulations are most probably associated with the two distinct types of neutron-star-forming supernova, with electron-capture supernovae preferentially producing systems with short spin periods, short orbital periods and low eccentricities. Intriguingly, the split between the two subpopulations is clearest in the distribution of the logarithm of spin period, a result that had not been predicted and which still remains to be explained.

Be/X-ray binaries (BeXs) are strong X-ray sources because their neutron stars accrete material at a relatively high rate. Their mass-losing Be-type companions are fast-rotating $8M_{\odot}-18M_{\odot}$ main-sequence stars that are surrounded by circumstellar 'decretion disks'. These disks are fuelled by the injection of mass and angular momentum at the stellar surface[6]. Neutron star spin periods in BeXs are typically 1 to 1,000 s, and BeX orbits are usually elliptical, with orbital periods ranging from about 10 to 1,000 d. Most of the accretion takes place during periastron passages, when the neutron star passes close to, or even through, the Be star decretion disk.

BeXs are exceptionally abundant in the Small Magellanic Cloud (SMC), where a burst of star formation about 60 Myr ago[7] seems to have produced a large population of these systems[8,9]. In fact, the SMC contains a comparable number of BeXs to the Milky Way, even though the mass ratio of the two galaxies is about 1:100. By contrast, the number of BeXs in the Large Magellanic Cloud (LMC) is broadly in line with its stellar mass content when compared with the Milky Way.

In the context of studying neutron star formation channels, it is useful to focus on well-defined, simple and 'clean' populations of neutron-star-hosting systems (that is, systems in which the orbital parameters have not yet evolved since the supernova) that nevertheless have a wide range of properties. BeXs can provide this. This is not only because the neutron stars in BeXs all have the same type of companion, but also because the accretion process itself seems to be universal, with the neutron star spin in or near an equilibrium state in which the magnetospheric radius of the neutron star equals the Keplerian co-rotation radius[10–12]. This conclusion is suggested empirically by the location of BeXs in the $\log(P_{\mathrm{orb}})$–$\log(P_{\mathrm{spin}})$ plane (the Corbet diagram[13]; P_{orb} and P_{spin} are the orbital and spin periods, respectively), where they tend to lie along a line with slope $\alpha \approx 2$ (Fig. 1).

The correlation in Fig. 1 between P_{spin} and P_{orb} among BeXs is highly significant, but the data have large scatter. Despite this scatter, however, the one-dimensional projections of the data onto the $\log(P_{\mathrm{orb}})$ and $\log(P_{\mathrm{spin}})$ axes both suggest that the BeX population might be bimodal. More specifically, the two subpopulations suggested by the data in Fig. 1 have characteristic periods of $P_{\mathrm{orb}} \approx 40$ d and $P_{\mathrm{spin}} \approx 10$ s (short-period mode) and $P_{\mathrm{orb}} \approx 100$ d and $P_{\mathrm{spin}} \approx 200$ s (long-period mode). The bimodality of the BeX population seems to be more prominent in $\log(P_{\mathrm{spin}})$ than in $\log(P_{\mathrm{orb}})$. This is helpful, because there are additional BeXs, not shown in Fig. 1, for which P_{spin} is known but P_{orb} is not. For the purpose of analysing the P_{spin} data on its own, we can therefore add these systems to the list of confirmed and probable BeXs.

Such an analysis is shown in Fig. 2. It confirms that the $\log(P_{\mathrm{spin}})$ distribution of BeXs contains two distinct subpopulations with characteristic spin periods of $P_{\mathrm{spin}} \approx 10$ s and $P_{\mathrm{spin}} \approx 200$ s and similar dispersions of about 0.4 dex. The short-P_{spin} and long-P_{spin} subpopulations contribute about 35 and 65% to the total number, respectively. The split into these two subpopulations is highly statistically significant for the full sample and remains significant even if the data set is divided by host galaxy. The split also remains significant if we consider only spectroscopically confirmed BeXs. Finally, the evidence for two subpopulations even remains significant if we use non-parametric statistical tests (which are less powerful, but more robust than the KMM algorithm; see the Supplementary Information for details).

As shown explicitly in Fig. 2d, the double-Gaussian decomposition of the independent SMC and Milky Way + LMC samples are consistent with each other. This makes it highly unlikely that selection effects are responsible for the observed bimodality. In any case, it is hard to conceive of a selection bias that would select specifically against BeXs with intermediate values of P_{spin} and/or P_{orb}. We therefore believe that the two modes of the $\log(P_{\mathrm{spin}})$ distribution correspond to physically distinct BeX subpopulations.

In principle, there are at least three ways to account for the existence of these subpopulations. First, they could correspond to two distinct neutron star spin equilibria that are accessible at all orbital periods. However, even though the bimodality is stronger in P_{spin} than in P_{orb}, the existence of the P_{spin}–P_{orb} correlation effectively rules out this possibility. Second, P_{orb} might be time dependent for BeXs, with the two subpopulations representing two distinct, long-lived evolutionary stages. However, the timescale for stellar-wind-driven changes in P_{orb} in the BeX phase, $\tau_{P_{\mathrm{orb}}} \approx 100$–1,000 Myr (refs 14,15), is substantially longer than the maximum duration of this phase, $\tau_{\mathrm{BeX,max}} \approx 20$ Myr, the lifetime of an $8M_{\odot}$ star. Thus, P_{orb} evolution also cannot account for the two observed subpopulations.

The third explanation that we consider is that the two subpopulations represent two distinct BeX formation channels. The most obvious possibility, with the farthest-reaching implications, is that the two channels are associated with the two distinct types of supernova event noted above. More specifically, an iron-core-collapse supernova marks the end point of the evolution of any sufficiently massive star, whereas an electron-capture supernova can occur only under highly restrictive

[1]University of Southampton, School of Physics and Astronomy, Southampton SO17 1BJ, UK. [2]University of Oxford, Department of Physics, Oxford OX1 3RH, UK.

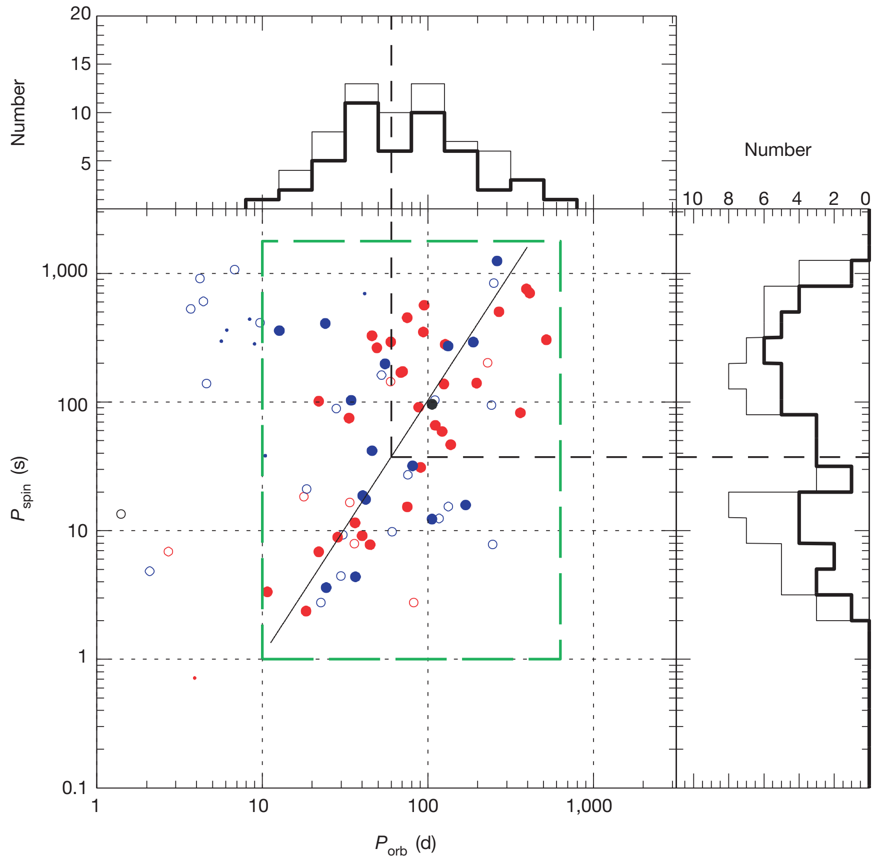


Figure 1 | The Corbet diagram for high-mass X-ray binaries. The central panel shows $\log(P_{orb})$ versus $\log(P_{spin})$ for neutron-star-hosting high-mass X-ray binaries. Filled circles correspond to spectroscopically confirmed BeXs, small dots to confirmed non-BeX systems and open circles to candidate BeXs. There are additional confirmed and candidate BeXs for which only P_{orb} or P_{spin} is known, but these are not shown. The dashed green lines mark a selection box that conservatively includes all confirmed BeXs for which P_{orb} and P_{spin} have been measured. Candidate systems outside this box are excluded from our sample of probable BeXs. The spin and orbital periods of confirmed and probable BeX systems are correlated. The Spearman rank correlation coefficient is $\rho = 0.49$ ($P = 3 \times 10^{-5}$, $N = 66$) for the full sample and $\rho = 0.49$ ($P = 4 \times 10^{-4}$, $N = 47$) for the confirmed systems (see Supplementary Information for a definition of P values). The scatter around the correlation is $\sigma_{\log(P_{spin})} = 0.7$ dex relative to the best-fitting line with slope $\alpha = 2$ (solid black line). Different colours indicate different host galaxies: blue, Milky Way; red, SMC; black, LMC. The histograms shown in the top and right-hand panels show the numbers of BeXs with spin and orbital periods in the respective ranges covered by the selection box. In each of these panels, the thick line corresponds to confirmed BeXs only, and the thin line corresponds to confirmed and probable BeXs. The vertical dashed line is drawn at $P_{orb} = 60$ d, the location of the apparent dip in the $\log(P_{orb})$ distribution. This value of P_{orb} corresponds to $P_{spin} \approx 40$ s (horizontal dashed line), which marks a more pronounced dip in the $\log(P_{spin})$ distribution.

conditions. In particular, an electron-capture supernova requires that the core reaches the critical density for electron capture to occur, $4.5 \times 10^9\ \mathrm{g\,cm^{-3}}$ (ref. 16). These conditions might be met in the late evolution of intermediate-mass stars[3,4] (those with initial masses satisfying $8M_\odot \lesssim M_{init} \lesssim 10M_\odot$), although the relevant mass range is uncertain and may be quite small[17,18]. However, it is much easier to meet the conditions for electron-capture supernovae naturally in binary systems[17].

The outcome of electron-capture supernovae differs from that of iron-core-collapse supernovae in two fundamental ways. First, electron-capture supernovae should produce somewhat less massive neutron stars ($\lesssim 1.3\ M_\odot$) than iron-core-collapse supernovae ($1.4\ M_\odot$) (ref. 3). Second, electron-capture supernovae are expected to impart much smaller kicks to the neutron stars they produce (average kick velocity of $\lesssim 50\ \mathrm{km\,s^{-1}}$) than are iron-core-collapse supernovae ($\gtrsim 200\ \mathrm{km\,s^{-1}}$) (ref. 17). In binary systems, where kicks induce orbital eccentricity, these differences could naturally give rise to two distinct subpopulations. The more conventional iron-core-collapse channel would produce high-eccentricity binaries containing high-mass neutron stars, and the electron-capture channel would produce low-eccentricity binaries containing low-mass neutron stars[16,17,19,20].

If this is the correct explanation for the two BeX populations we have discovered, they should differ not only in P_{spin} and P_{orb}, but also in their characteristic orbital eccentricities, e. These have so far been measured for only about 20 BeXs. Figure 3 shows the distribution of these systems in the $\log(P_{spin})$–e plane. Even though there are only eight such BeXs with $P_{spin} \gtrsim 40$ s, it seems that long-P_{spin} systems have preferentially higher eccentricities than short-P_{spin} systems. The figure also shows the cumulative eccentricity distributions of the slow ($P_{spin} < 40$ s) and fast ($P_{spin} > 40$ s) BeX pulsar subpopulations. A Kolmogorov–Smirnov test shows that, despite the small number of systems for which eccentricity has been measured, the maximum difference between these distributions is marginally significant (Supplementary Information).

The two BeX populations that we have discovered are more clearly separated in P_{spin} than in P_{orb}. Given that P_{orb} does not evolve significantly within the BeX phase, whereas P_{spin} does, P_{orb} might be expected to be the more faithful tracer of the formation channels. However, the P_{orb} distribution after the supernova must depend strongly on the P_{orb} distribution before the supernova, and both low- and high-velocity kicks can in principle produce a wide range of post-supernova orbital periods[17]. This may explain why the bimodality is only marginally observed in the P_{orb} distribution. By contrast, the equilibrium spin period is expected to depend on several system parameters other than P_{orb} (refs 10–12). If any of these parameters differ systematically between BeXs produced by the iron-core-collapse and electron-capture

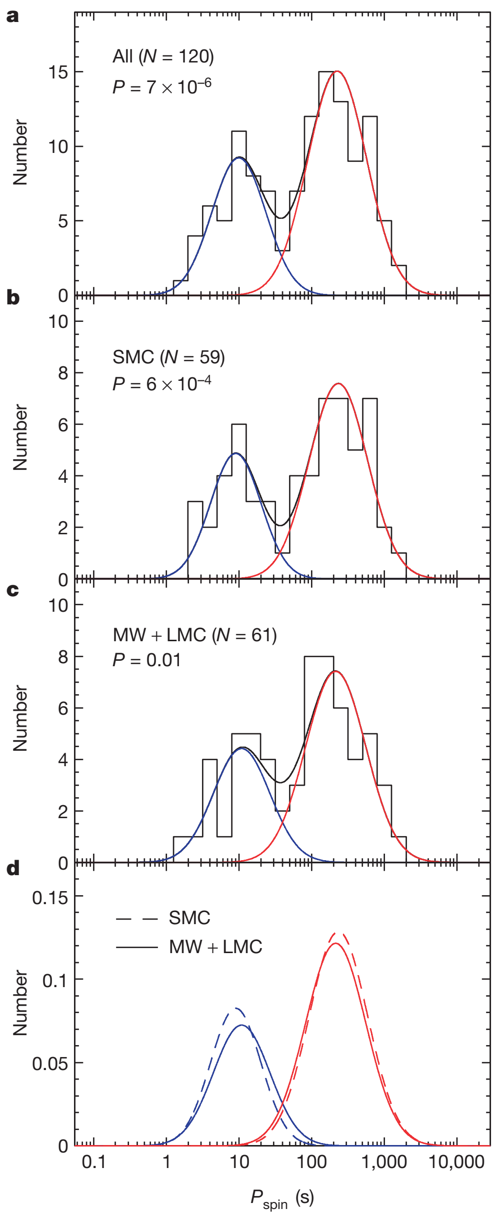


Figure 2 | The log(P_{spin}) distribution of confirmed and probable BeXs. **a**, Distribution for all systems. **b**, **c**, Distribution broken down by host galaxy: SMC (**b**); Milky Way (MW) + LMC (**c**). All of these distributions are bimodal (modes shown in red and blue), and the double-Gaussian decomposition (black) suggested by the KMM algorithm[25] (an algorithm that detects bimodality in an observational data set) is shown in each panel. The number of systems contributing to each observed distribution and the associated P value provided by the algorithm are also shown. Applying the KMM test to the subset of spectroscopically confirmed systems (not shown) gives $P = 8 \times 10^{-3}$ ($N = 64$). **d**, Direct comparison of the decompositions for the independent SMC and Milky Way + LMC populations, showing them to be mutually consistent. Additional details regarding the statistical evidence for the existence of distinct subpopulations in the P_{spin} data are given in Supplementary Information.

channels, P_{spin} may be a more reliable indicator of formation channel than P_{orb}.

Our results suggest numerous avenues for further research. First, it is important to expand the database of BeXs with reliable measurements of P_{spin}, P_{orb} and e, to confirm and further quantify our findings. Second, if short-P_{spin} BeXs are formed by low-kick-velocity electron-capture supernovae, they should have systematically smaller space velocities than long-P_{spin} BeXs. This prediction might be testable[21,22]. Third, short-P_{spin} systems should also have systematically lower neutron star masses than long-P_{spin} systems[16,17,19,20]. This prediction might also be testable[23]. Fourth, although our discovery of two populations of BeX pulsars is robust, our suggested explanation for their origin is clearly speculative and demands a fuller investigation. Intriguingly, recent binary population synthesis calculations have shown that the electron-capture supernova channel may be very efficient at forming BeXs[24]. However, whereas the same population synthesis models also suggest that the electron-capture channel accounts for the overabundance of BeXs in the SMC[24], the two BeX populations that we have discovered seem to have similar relative abundances in the SMC and the Milky Way.

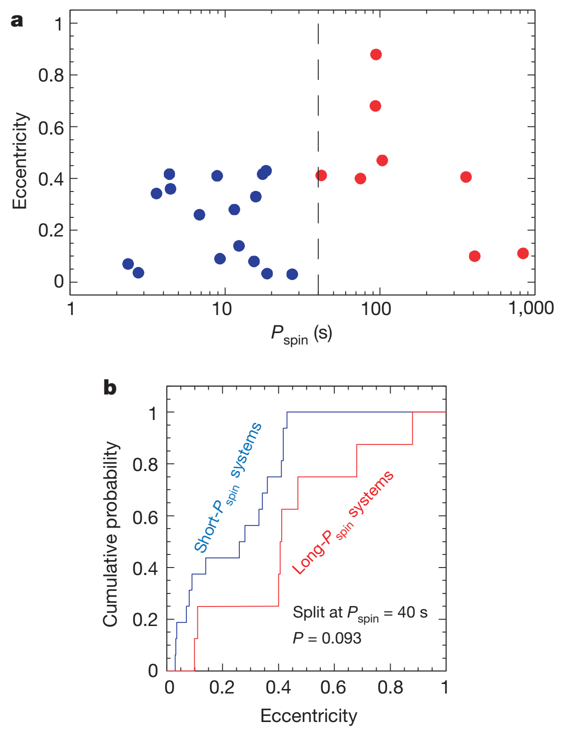


Figure 3 | The dependence of eccentricity on P_{spin} among BeXs. **a**, P_{spin} versus eccentricity for all confirmed and probable BeXs with measured spin periods and eccentricities. The vertical dashed line marks the approximate division between the short-P_{spin} and long-P_{spin} subpopulations (Figs 1 and 2). **b**, Cumulative eccentricity distributions of these two subpopulations. A Kolmogorov–Smirnov test provides marginal evidence for a difference between these distributions ($P = 0.093$), with the short-P_{spin} population being characterized by lower eccentricities (see Supplementary Information for additional discussion).

Received 5 May; accepted 2 September 2011.

Published online 9 November 2011.

1. Heger, A., Fryer, C. L., Woosley, S. E., Langer, N. & Hartmann, D. H. How massive single stars end their life. *Astrophys. J.* **591,** 288–300 (2003).
2. Woosley, S. & Janka, T. The physics of core-collapse supernovae. *Nature Phys.* **1,** 147–154 (2005).
3. Nomoto, K. Evolution of 8–10 solar mass stars toward electron capture supernovae. I - Formation of electron-degenerate O + NE + MG cores. *Astrophys. J.* **277,** 791–805 (1984).
4. Nomoto, K. Evolution of 8–10 solar mass stars toward electron capture supernovae. II - Collapse of an O + NE + MG core. *Astrophys. J.* **322,** 206–214 (1987).
5. Reig, P. Be/X-ray binaries. *Astrophys. Space Sci.* **332,** 1–29 (2011).
6. Lee, U., Osaki, Y. & Saio, H. Viscous excretion discs around Be stars. *Mon. Not. R. Astron. Soc.* **250,** 432–437 (1991).
7. Harris, J. & Zaritsky, D. The star formation history of the Small Magellanic Cloud. *Astron. J.* **127,** 1531–1544 (2004).
8. Haberl, F. & Pietsch, W. X-ray observations of Be/X-ray binaries in the SMC. *Astron. Astrophys.* **414,** 667–676 (2004).
9. Coe, M. J., Edge, W. R. T., Galache, J. L. & McBride, V. A. Optical properties of Small Magellanic Cloud X-ray binaries. *Mon. Not. R. Astron. Soc.* **356,** 502–514 (2005).
10. Waters, L. B. F. M. & van Kerkwijk, M. H. The relation between orbital and spin periods in massive X-ray binaries. *Astron. Astrophys.* **223,** 196–206 (1989).
11. Li, X. & van den Heuvel, E. P. J. On the relation between spin and orbital periods in Be/X-ray binaries. *Astron. Astrophys.* **314,** L13–L16 (1996).

12. Liu, Q. Z., Li, X. D. & Wei, D. M. in *High Energy Processes and Phenomena in Astrophysics* (eds Li, X. D., Trimble, V. & Wang, Z. R.) 215–217 (Proc. IAU Symp. 214, Univ. Chicago Press, 2003).
13. Corbet, R. H. D. Be/neutron star binaries: a relationship between orbital period and neutron star spin period. *Astron. Astrophys.* **141,** 91–93 (1984).
14. de Jager, C., Nieuwenhuijzen, H. & van der Hucht, K. A. Mass loss rates in the Hertzsprung-Russell diagram. *Astron. Astrophys. Suppl. Ser.* **72,** 259–289 (1988).
15. Tout, C. A. & Hall, D. S. Wind driven mass transfer in interacting binary systems. *Mon. Not. R. Astron. Soc.* **253,** 9–18 (1991).
16. Podsiadlowski, P. *et al.* The double pulsar J0737–3039: testing the neutron star equation of state. *Mon. Not. R. Astron. Soc.* **361,** 1243–1249 (2005).
17. Podsiadlowski, P. *et al.* The effects of binary evolution on the dynamics of core collapse and neutron star kicks. *Astrophys. J.* **612,** 1044–1051 (2004).
18. Poelarends, A. J. T., Herwig, F., Langer, N. & Heger, A. The supernova channel of super-AGB stars. *Astrophys. J.* **675,** 614–625 (2008).
19. van den Heuvel, E. P. J. in *ESA SP-552: Proc. 5th INTEGRAL Workshop ''The INTEGRAL Universe''* (eds Schoenfelder, V., Lichti, G. & Winkler, C.) 185–194 (ESA Spec. Publ. 552, European Space Agency, 2004).
20. Schwab, J., Podsiadlowski, P. & Rappaport, S. Further evidence for the bimodal distribution of neutron-star masses. *Astrophys. J.* **719,** 722–727 (2010).
21. Coe, M. J. An estimate of the supernova kick velocities for high-mass X-ray binaries in the Small Magellanic Cloud. *Mon. Not. R. Astron. Soc.* **358,** 1379–1382 (2005).
22. Antoniou, V., Zezas, A., Hatzidimitriou, D. & Kalogera, V. Star formation history and X-ray binary populations: the case of the Small Magellanic Cloud. *Astrophys. J.* **716,** L140–L145 (2010).
23. Coe, M. J., McBride, V. A. & Corbet, R. H. D. Exploring accretion theory with X-ray binaries in the Small Magellanic Cloud. *Mon. Not. R. Astron. Soc.* **401,** 252–256 (2010).
24. Linden, T., Sepinsky, J. F., Kalogera, V. & Belczynski, K. Probing electron-capture supernovae: X-Ray binaries in starbursts. *Astrophys. J.* **699,** 1573–1577 (2009).
25. Ashman, K. M., Bird, C. M. & Zepf, S. E. Detecting bimodality in astronomical datasets. *Astron. J.* **108,** 2348–2361 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Research support for this project was provided by the UK Science and Technology Facilities Council. We would like to thank T. Maccarone and T. Linden for discussions.

Author Contributions C.K. carried out the statistical analysis for this project and wrote most of the text. M.J.C. compiled the high-mass X-ray binary data set that forms the basis for our analysis, and collaborated with C.K. on all aspects of the project from its inception. P.P. contributed to the theoretical interpretation of the results and to the final text. All authors discussed the results and their presentation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.K. (c.knigge@soton.ac.uk).

doi:10.1038/nature10561

Observation of the dynamical Casimir effect in a superconducting circuit

C. M. Wilson[1], G. Johansson[1], A. Pourkabirian[1], M. Simoen[1], J. R. Johansson[2], T. Duty[3], F. Nori[2,4] & P. Delsing[1]

One of the most surprising predictions of modern quantum theory is that the vacuum of space is not empty. In fact, quantum theory predicts that it teems with virtual particles flitting in and out of existence. Although initially a curiosity, it was quickly realized that these vacuum fluctuations had measurable consequences—for instance, producing the Lamb shift[1] of atomic spectra and modifying the magnetic moment of the electron[2]. This type of renormalization due to vacuum fluctuations is now central to our understanding of nature. However, these effects provide indirect evidence for the existence of vacuum fluctuations. From early on, it was discussed whether it might be possible to more directly observe the virtual particles that compose the quantum vacuum. Forty years ago, it was suggested[3] that a mirror undergoing relativistic motion could convert virtual photons into directly observable real photons. The phenomenon, later termed the dynamical Casimir effect[4,5], has not been demonstrated previously. Here we observe the dynamical Casimir effect in a superconducting circuit consisting of a coplanar transmission line with a tunable electrical length. The rate of change of the electrical length can be made very fast (a substantial fraction of the speed of light) by modulating the inductance of a superconducting quantum interference device at high frequencies (>10 gigahertz). In addition to observing the creation of real photons, we detect two-mode squeezing in the emitted radiation, which is a signature of the quantum character of the generation process.

That mirrors can be used to measure vacuum fluctuations was first predicted by Casimir[6] in 1948. Casimir predicted that two mirrors, that is, perfectly conducting metal plates, held parallel to each other in vacuum will experience an attractive force. Essentially, the mirrors reduce the density of electromagnetic modes between them. The vacuum radiation pressure between the plates is then less than the pressure outside, generating the force. As this static Casimir effect can then be explained by a mismatch of vacuum modes in space, the dynamical Casimir effect can be seen as arising from a mismatch of vacuum modes in time. As a mirror moves, it changes the spatial mode structure of the vacuum. If the mirror's velocity, v, is slow compared to the speed of light, c, the electromagnetic field can adiabatically adapt to the changes and no excitation occurs. If instead v/c is not negligible, then the field cannot adjust smoothly and can be non-adiabatically excited out of the vacuum.

The static Casimir effect can also be calculated in terms of the electrical response of the mirrors to the electromagnetic field[7]. A similar complementary explanation exists for the dynamical Casimir effect[3]. An ideal mirror represents a boundary condition for the electromagnetic field—in particular, that the electric field is zero at the surface. This boundary condition is enforced by the flow of screening currents in the metal. A mirror moving in a finite electromagnetic field then loses energy, because the screening currents will emit electromagnetic radiation, as in an antenna. Classically, we expect this radiation damping to be zero in a region where the electric field strength is zero. In quantum theory, however, vacuum fluctuations will always generate screening currents. Therefore, even moving in the vacuum can cause a mirror to emit real photons in response to vacuum fluctuations.

If we consider the real experiment of moving a physical mirror near the speed of light, we quickly see that it is not feasible. This fact has led to a number of alternative proposals[8–20], for instance using surface acoustic waves, nanomechanical resonators, or modulation of the electrical properties of a cavity.

Here we investigate one such proposal using a superconducting circuit[16,17]: an open transmission line terminated by a SQUID (superconducting quantum interference device). A SQUID is composed of two Josephson junctions connected in parallel to form a loop. At the frequencies studied here, the SQUID acts as a parametric inductor whose value, L_J, can be tuned by applying a magnetic flux, Φ_{ext}, through the SQUID loop. When placed at the end of a transmission line, this SQUID can then be used to change the line's boundary condition. In previous work[18,21], we showed that this tuning can be done on very short timescales. The changing inductance can be described as a change in the electrical length of the transmission line and, in fact, provides the same time-dependent boundary condition as the idealized moving mirror[22,23]. In the same way as for the mirror, the boundary condition is enforced by screening currents that flow through the SQUID. Unlike the mirror, the maximum effective velocity of the boundary, v_e, defined as the rate of change of the electrical length, can be very large compared to the speed of light in the transmission line, $c_0 \approx 0.4c$, approaching $v_e/c_0 \approx 0.25$ for large modulations of L_J. The photon production rate is therefore predicted to be several orders of magnitude larger than in other systems[17].

Quantum theory allows us to make more detailed predictions than just that photons will be produced. If the boundary is driven sinusoidally at an angular frequency $\omega_d = 2\pi f_d$, then it is predicted[17,24] that photons will be produced in pairs such that their frequencies, ω_+ and ω_-, sum to the drive frequency, that is, we expect $\omega_d = \omega_+ + \omega_-$. This pairwise production implies that the electromagnetic field at these sideband frequencies, symmetric around $\omega_d/2$, should be correlated. In detail, we can predict that the field should exhibit what is known as two-mode squeezing (TMS)[25].

Theoretically, we treat the problem as a scattering problem in the context of quantum network theory[26] (see Supplementary Information). If the boundary is driven with a small amplitude $\delta\ell_e$, we find that the output photon flux density at frequency ω for an input thermal state is

$$n_{out}(\omega) = n_{in}(\omega) + |S(\omega)|^2 n_{in}(\omega_d - \omega) + |S(\omega)|^2 \qquad (1)$$

where $S(\omega) = -i(\delta\ell_e/c_0)\sqrt{\omega(\omega_d-\omega)}A(\omega)A^*(\omega_d-\omega)$, $A(\omega)$ is the spectral amplitude of the transmission line and $A^*(\omega)$ is its complex conjugate. The first two terms on the right-hand side of equation (1), proportional to the thermal occupation number $n_{in}(\omega)$, represent the purely classical effects of reflection and upconversion of the input field to the drive frequency. They are zero at zero temperature. The last term on the right-hand side is due to vacuum fluctuations and is, in fact, the DCE radiation.

[1]Department of Microtechnology and Nanoscience, Chalmers University of Technology, Göteborg 412 96, Sweden. [2]Advanced Science Institute, RIKEN, Wako-shi, Saitama 351-0198, Japan. [3]University of New South Wales, Sydney, New South Wales 2052, Australia. [4]University of Michigan, Ann Arbor, Michigan 48109, USA.

The photon production rate depends on the density of states in the transmission line, which is $|A(\omega)|^2$. For an ideal transmission line, $A(\omega)=1$ and the DCE radiation measured at a detuning $\delta\omega$ from $\omega_d/2$ becomes $n_{\text{out}}^{\text{DCE}}(\varepsilon)=(v_e/2c_0)^2(1-\varepsilon^2)$ where $\varepsilon=2\delta\omega/\omega_d$ is the normalized detuning and $v_e=\delta\ell_e\omega_d$. The integrated photon flux of the DCE radiation is then $\Gamma_{\text{DCE}}=(\omega_d/12\pi)(v_e/c_0)^2$, identical to that of an ideal mirror oscillating in one-dimensional space[23]. The relativistic nature of the effect is apparent here, in that the photon flux goes to zero if we allow the speed of light to go to infinity.

We present measurements on two samples, both of which consist of a SQUID connected to an aluminium coplanar waveguide (CPW). Sample 1 has a long (~43 mm) CPW and sample 2 has a short (~0.1 mm) CPW (Fig. 1a, b). The samples are cooled to ≤50 mK in a dilution refrigerator. This corresponds to a thermal photon occupation number of $n_{\text{in}}<0.01$ at 5 GHz, which is the centre of our analysis band. In separate measurements of qubit circuits in the same set-up, we have demonstrated[27] that the radiation temperature of the system is near the cryostat's base temperature. If we consider the last

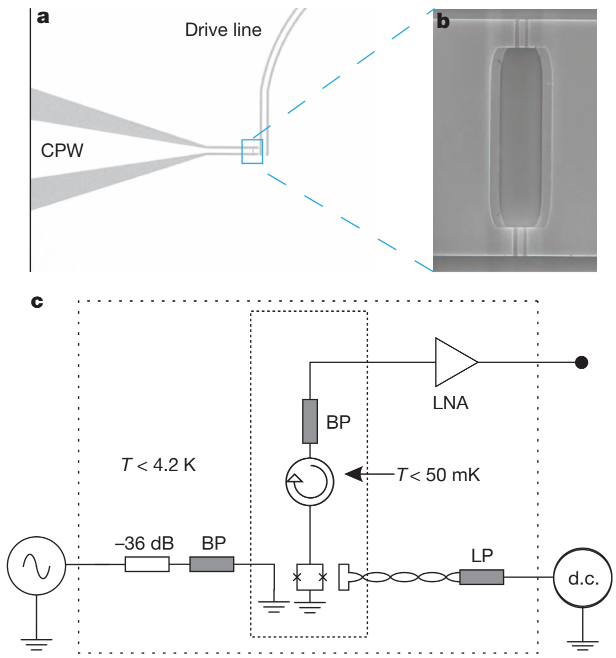


Figure 1 | Experimental overview. **a**, Optical micrograph of sample 2. Light parts are Al, which fills most of the image, while the dark parts are the Si substrate, visible where the Al has been removed to define the transmission lines. The output line is labelled CPW and the drive line enters from the top. Both lines converge near the SQUID (boxed). **b**, A scanning-electron micrograph of the SQUID. The SQUID has a vertical dimension of 13 μm and a normal state resistance of 218 Ω (170 Ω) implying $L_J(0)=0.23$ nH (0.18 nH) for sample 2 (sample 1). A basic electrical characterization of the SQUID is presented in Supplementary Fig. 1. **c**, A simplified schematic of the measurement set-up. The SQUID is indicated by the box with two crosses, suggestive of the SQUID loop interrupted by Josephson junctions. A small external coil is also used to apply a d.c. flux bias through a lowpass filter (LP). The driving line has 36 dB of cold attenuation, along with an 8.4–12 GHz bandpass filter (BP). The filter ensures that no thermal radiation couples to the transmission line in the frequency region were we expect DCE radiation. (For sample 1, the last 6 dB of attenuation were at base temperature.) The outgoing field of the CPW is coupled through two circulators to a cryogenic low-noise amplifier (LNA) with a system noise temperature of $T_N\approx 6$ K. At room temperature, the signal is further amplified before being captured by two vector microwave digitizers. The dashed boxes delineate portions of the set-up at different temperatures, T, which are labelled.

two terms on the right-hand side of equation (1), which are the response of the system to the changing boundary, we can compare this small value $n_{\text{in}}=0.01$ to the vacuum response, which has a coefficient of 1.

To study the effects of non-adiabatic perturbations, we drive the flux through the SQUID at microwave frequencies using an inductively coupled CPW line that is short-circuited ~20 μm from the SQUID. We measure the output power from the measurement line as a function of drive power and frequency, $f_d=\omega_d/2\pi$. We start with the analysis frequency tracking the drive at $f_d/2$, where we expect the DCE radiation to be centred. The results are shown in Supplementary Fig. 2. In both samples, we clearly see photon generation for essentially all drive frequencies spanning the 8–12 GHz band set by the filtering of the line. This corresponds to an analysis band of 4–6 GHz.

In the next set of measurements, we fix the drive frequency, but scan the analysis frequency. In this way, we can see over what band photons are produced for fixed drive frequency. In both samples, we clearly see broadband photon production for all the frequencies analysed, including detunings from $f_d/2$ larger than 2 GHz. In Fig. 2a and b, we show results for sample 1. The broadband nature of the photon generation clearly distinguishes the observed phenomenon from that of a parametric amplifier, which has a narrow band defined by a resonator.

We quantify the photon production rate by referencing it to our system noise temperature of $T_N\approx 6$ K, which has been calibrated previously using a shot-noise thermometer[28]. The measured power spectral density is corrected for the transmission variations measured with a network analyser and then divided by the photon energy, $\hbar\omega$, at each analysis frequency to convert to n_{out}. In Fig. 2a and b, we clearly see a corrugated structure in n_{out}, reflecting variations in $A(\omega)$ caused by parasitic reflections in the measurement line at, for example, cable connectors or the input of the amplifier. Overall, we see that n_{out} is clearly symmetric around $f_d/2$, as emphasized in Fig. 2c and d, which strongly indicates that the radiation is produced by the DCE and not by a spurious effect, such as heating.

To make a quantitative comparison to theory, we need to know the drive amplitude, which is difficult to know a priori as the short-circuit termination of the pump line is not matched to the characteristic impedance of the line. We therefore expect the current flowing through the line to be a strong function of frequency. However, we see that n_{out} starts to saturate as a function of drive power above the nominal level of 100 pW (not shown). We now make the reasonable assumption that this saturation is caused by the drive amplitude reaching the level where the modulation of L_J saturates, that is, where the sum of the flux amplitude and d.c. bias flux reaches $0.5\Phi_0$, where $\Phi_0=h/2e$ is the superconducting flux quantum. For our working point of $\Phi_{\text{ext}}=-0.35\Phi_0$, the corresponding flux amplitude is only $0.15\Phi_0$, which we have previously demonstrated is experimentally accessible[21]. For these large modulations, equation (1) is no longer sufficient because higher-order processes become important. However, these processes can be included and n_{out} computed numerically[17]. As shown in Fig. 2e, we find that the calculated n_{out} matches the measured value if we assume modest values of $|A(\omega_+)A(\omega_-)|\approx 2$–$4$. This corresponds to a voltage standing wave ratio (VSWR) in our measurement line of less than 2, which is reasonable. We note that at these large drives, n_{out} deviates from its linear dependence on drive power, which is proportional to $(v_e/c_0)^2$. However, we see in Fig. 2e that the full theory[17] captures this deviation well. We also comment that, in making the calculation, we have made a model of the electromagnetic environment, encapsulated in $A(\omega)$, to high frequencies. There is a great deal of uncertainty in doing this. However, with a basic physical model, we see that we are able to reproduce both the magnitude and drive dependence of n_{out}.

Theory[17] also predicts that the output should exhibit voltage–voltage correlations, known as TMS. Experimentally, we measure the four quadrature voltages of the upper and lower sidebands, corresponding

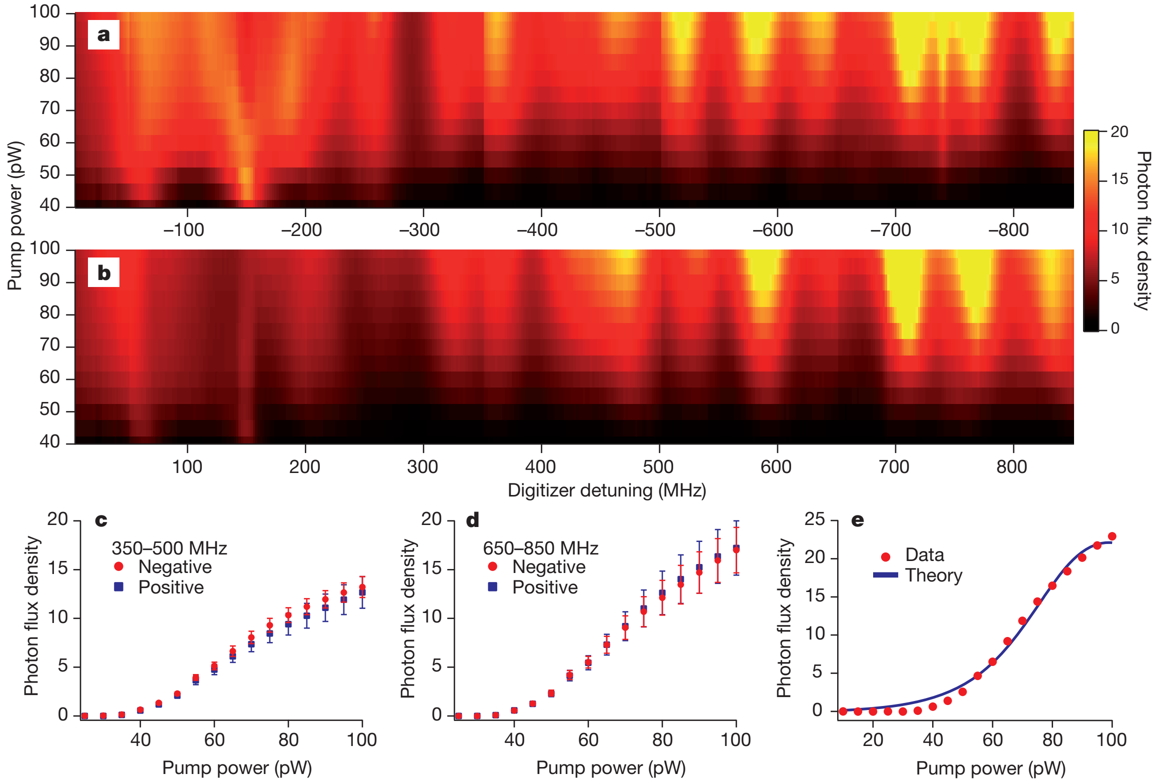
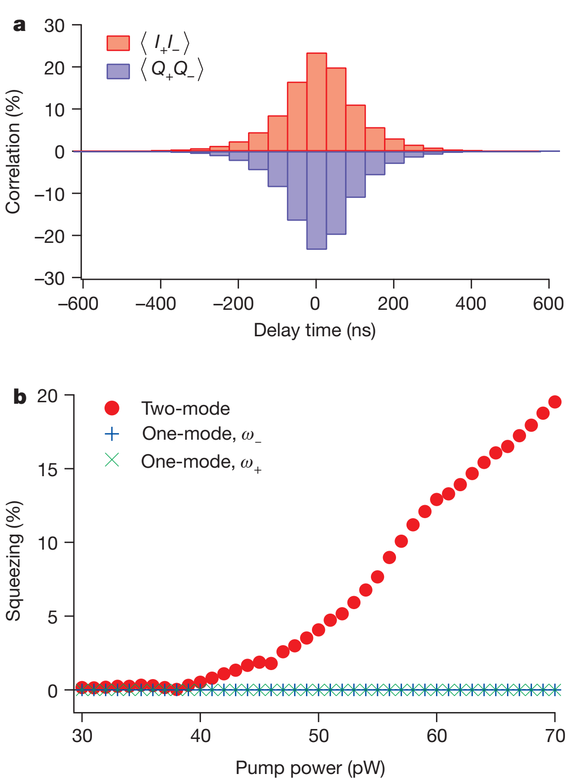


Figure 2 | **Photons generated by the dynamical Casimir effect.** Here we show the output flux of the transmission line while driving sample 1 at $f_d = 10.30$ GHz. **a**, **b**, Broadband photon generation. We plot the dimensionless photon flux density, n_{out} (photons s^{-1} Hz^{-1}), which is the measured power spectral density normalized to the photon energy, $\hbar\omega$, as a function of pump power and detuning, $\delta\omega/2\pi$. Panel **a** shows negative detunings (axis reversed), while **b** shows positive detunings. The symmetry of the spectrum is apparent. Positive and negative detunings are recorded simultaneously. The plots are stitched together from several separate scans, between which we have changed image rejection filters at the input of the analysers. **c**, **d**, The photon flux density for positive and negative detunings averaged over frequency (at fixed power) for two different symmetric bands, showing the symmetry of the spectrum. Error bars, s.d. **e**, A section through **a** at $\delta\omega/2\pi = -764$ MHz, along with a fit to the full theory of ref. 17.



to the observable (Hermitian) quadrature operators $I_\pm$ and $Q_\pm$. We can write the TMS, σ_2, in terms of the quadratures as

$$\sigma_2 = \frac{1}{P_{avg}}\left(\langle I_+I_-\rangle - \langle Q_+Q_-\rangle\right) \qquad (2)$$

where $P_{avg} = \left(\langle I_+^2\rangle + \langle Q_+^2\rangle + \langle I_-^2\rangle + \langle Q_-^2\rangle\right)/2$ is the average noise power in the sidebands and $\langle ...\rangle$ denotes the expectation value (Supplementary Information). We also expect a special structure for the correlations, in particular that $\langle I_+I_-\rangle = -\langle Q_+Q_-\rangle$ and that $\langle I_+Q_-\rangle = \langle I_-Q_+\rangle$. Finally, we comment[25] that by the proper choice of analysis phase, we can specify $\langle I_+Q_-\rangle = \langle I_-Q_+\rangle = 0$ without loss of generality, which has been done in writing equation (2) (see Supplementary Information and Supplementary Fig. 4).

To measure the correlations, we use a single amplifier but take advantage of the fact that the amplifier noise at different frequencies is uncorrelated. After amplifying, we split the signal into two separate analysis chains. We then calculate the four time-averaged *IQ* cross-correlation functions. Typical results are shown in Fig. 3. We see very clear cross-correlations that are ~1,000 times larger than the parasitic amplifier correlation (see Supplementary Fig. 3). Also, we see that

Figure 3 | **Two-mode squeezing of the DCE field.** **a**, The normalized cross-correlation functions $\langle I_+I_-\rangle/P_{avg}$ and $\langle Q_+Q_-\rangle/P_{avg}$, measured on sample 1, using $f_d = 10.30$ GHz and $|\delta\omega/2\pi| = 833$ MHz. We clearly see cross-correlations of the order of 25%, and that $\langle I_+I_-\rangle = -\langle Q_+Q_-\rangle$, as predicted. The shape of the correlation functions in delay time is determined by the filtering of the time traces. **b**, The two-mode squeezing, σ_2, of the field along with the one-mode squeezing, σ_1, at both ω_+ and ω_- as a function of drive power, measured on sample 1 at $|\delta\omega/2\pi| = 588$ MHz. We see that σ_2 clearly increases while the single-mode fields remain unsqueezed.

indeed $\langle I_+ I_- \rangle = -\langle Q_+ Q_- \rangle$, as we expect for TMS. The correlations imply a value of $\sigma_2 \approx 0.46$, which compares well to the predicted maximum squeezing of 50%.

Theory further predicts[25] that, even though the field is two-mode squeezed, if we look at either sideband frequency individually, it will remain unsqueezed, essentially appearing as a thermal field at some effective temperature. In Fig. 3b, we plot the TMS of the field, σ_2, along with the one-mode squeezing, $\sigma_1 = (\langle I^2 \rangle - \langle Q^2 \rangle)/(\langle I^2 \rangle + \langle Q^2 \rangle)$ at both frequencies, ω_+ and ω_-, as a function of drive power. We clearly see that σ_2 increases as a function of drive power while the one-mode fields remain unsqueezed.

In the Supplementary Discussion, we consider, and rule out, a number of spurious effects that could be the source of n_{out}. However, even if we assume that the photon creation is connected to the non-adiabatic modulation of the boundary condition, we need to confirm that it is seeded by vacuum fluctuations, not by spurious noise in the measurement system. To check this, we measured n_{out} with the cryostat temperature elevated to 250 mK, which is roughly $\hbar\omega_d/2k_B$. The direct comparison of the fluxes is complicated by the fact the aluminium CPW becomes lossy at this temperature, so some power will be lost. Still, we measure the ratio of the output fluxes to be 1.4 ± 0.3, which agrees with the expected value of 1.6 assuming the starting temperature is 50 mK. This tells us that $n_{\mathrm{in}} \ll 1$ at the base temperature, and that our system is therefore dominated by vacuum effects.

Received 31 August; accepted 15 September 2011.

1. Scully, M. O. & Zubairy, M. S. *Quantum Optics* (Cambridge Univ. Press, 1997).
2. Greiner, W. & Schramm, S. Resource letter QEDV-1: the QED vacuum. *Am. J. Phys.* **76,** 509–518 (2008).
3. Moore, G. Quantum theory of the electromagnetic field in a variable-length one-dimensional cavity. *J. Math. Phys.* **11,** 2679–2691 (1970).
4. Dodonov, V. Current status of the dynamical Casimir effect. *Phys. Scripta* **82,** 038105 (2010).
5. Dalvit, D. A. R., Neto, P. A. M. & Mazzitelli, F. D. Fluctuations, dissipation and the dynamical Casimir effect. Preprint at ⟨http://arXiv.org/abs/1006.4790v2⟩ (2010).
6. Casimir, H. B. G. On the attraction between two perfectly conducting plates. *Proc. K. Ned. Akad. Wet. B* **51,** 793 (1948).
7. Lamoreaux, S. K. Casimir forces: still surprising after 60 years. *Phys. Today* **60,** 40–45 (2007).
8. Braggio, C. *et al.* A novel experimental approach for the detection of the dynamical Casimir effect. *Europhys. Lett.* **70,** 754–760 (2005).
9. Yablonovitch, E. Accelerating reference frame for electromagnetic waves in a rapidly growing plasma: Unruh-Davies-Fulling-Dewitt radiation and the nonadiabatic Casimir effect. *Phys. Rev. Lett.* **62,** 1742–1745 (1989).
10. Lozovik, Y., Tsvetus, V. & Vinogradov, E. Femtosecond parametric excitation of electromagnetic field in a cavity. *JETP Lett.* **61,** 723–729 (1995).
11. Dodonov, V., Klimov, A. & Nikonov, D. Quantum phenomena in nonstationary media. *Phys. Rev. A* **47,** 4422–4429 (1993).
12. Schützhold, R., Plunien, G. & Soff, G. Quantum radiation in external background fields. *Phys. Rev. A* **58,** 1783–1793 (1998).
13. Kim, W., Brownell, J. & Onofrio, R. Detectability of dissipative motion in quantum vacuum via superradiance. *Phys. Rev. Lett.* **96,** 200402 (2006).
14. Liberato, S., Ciuti, C. & Carusotto, I. Quantum vacuum radiation spectra from a semiconductor microcavity with a time-modulated vacuum Rabi frequency. *Phys. Rev. Lett.* **98,** 103602 (2007).
15. Günter, G. *et al.* Sub-cycle switch-on of ultrastrong light–matter interaction. *Nature* **458,** 178–181 (2009).
16. Johansson, J. R., Johansson, G., Wilson, C. M. & Nori, F. Dynamical Casimir effect in a superconducting coplanar waveguide. *Phys. Rev. Lett.* **103,** 147003 (2009).
17. Johansson, J. R., Johansson, G., Wilson, C. M. & Nori, F. Dynamical Casimir effect in superconducting microwave circuits. *Phys. Rev. A* **82,** 052509 (2010).
18. Wilson, C. M. *et al.* Photon generation in an electromagnetic cavity with a time-dependent boundary. *Phys. Rev. Lett.* **105,** 233907 (2010).
19. Dezael, F. & Lambrecht, A. Analogue Casimir radiation using an optical parametric oscillator. *Europhys. Lett.* **89,** 14001 (2010).
20. Nation, P. D., Johansson, J. R., Blencowe, M. P. & Nori, F. Stimulating uncertainty: amplifying the quantum vacuum with superconducting circuits. *Rev. Mod. Phys.* (in the press); preprint at ⟨http://arXiv.org/abs/1103.0835v1⟩ (2011).
21. Sandberg, M. *et al.* Tuning the field in a microwave resonator faster than the photon lifetime. *Appl. Phys. Lett.* **92,** 203501 (2008).
22. Fulling, S. A. & Davies, P. C. W. Radiation from a moving mirror in two dimensional space-time: conformal anomaly. *Proc. R. Soc. Lond. A* **348,** 393–414 (1976).
23. Lambrecht, A., Jaekel, M. & Reynaud, S. Motion induced radiation from a vibrating cavity. *Phys. Rev. Lett.* **77,** 615–618 (1996).
24. Dodonov, V., Klimov, A. & Man'ko, V. Generation of squeezed states in a resonator with a moving wall. *Phys. Lett. A* **149,** 225–228 (1990).
25. Caves, C. M. & Schumaker, B. L. New formalism for 2-photon quantum optics. 1. Quadrature phases and squeezed states. *Phys. Rev. A* **31,** 3068–3092 (1985).
26. Yurke, B. & Denker, J. S. Quantum network theory. *Phys. Rev. A* **29,** 1419–1437 (1984).
27. Hoi, I.-C. *et al.* Demonstration of a single-photon router in the microwave regime. *Phys. Rev. Lett.* **107,** 073601 (2011).
28. Spietz, L., Lehnert, K., Siddiqi, I. & Schoelkopf, R. Primary electronic thermometry using the shot noise of a tunnel junction. *Science* **300,** 1929–1932 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Milburn and V. Shumeiko for discussions, and J. Aumentado and L. Spietz for providing the shot-noise thermometer. C.M.W., P.D., G.J., A.P. and M.S. were supported by the Swedish Research Council, the Wallenberg Foundation, STINT and the European Research Council. F.N. and J.R.J. acknowledge partial support from the LPS, NSA, ARO, DARPA, AFOSR, NSF grant no. 0726909, Grant-in-Aid for Scientific Research (S), MEXT Kakenhi on Quantum Cybernetics, and the JSPS-FIRST programme. T.D. acknowledges support from STINT and the Australian Research Council (grants DP0986932 and FT100100025).

Author Contributions The experimental work was carried out by C.M.W., A.P., M.S., T.D. and P.D. The theoretical work was performed by J.R.J., F.N., C.M.W. and G.J.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.M.W. (chris.wilson@chalmers.se).

LETTER

doi:10.1038/nature10593

Atom-resolved imaging of ordered defect superstructures at individual grain boundaries

Zhongchang Wang[1], Mitsuhiro Saito[1], Keith P. McKenna[1,2], Lin Gu[1], Susumu Tsukimoto[1], Alexander L. Shluger[1,2] & Yuichi Ikuhara[1,3,4]

The ability to resolve spatially and identify chemically atoms in defects would greatly advance our understanding of the correlation between structure and property in materials[1]. This is particularly important in polycrystalline materials, in which the grain boundaries have profound implications for the properties and applications of the final material[2]. However, such atomic resolution is still extremely difficult to achieve, partly because grain boundaries are effective sinks for atomic defects and impurities[3–5], which may drive structural transformation of grain boundaries and consequently modify material properties[6,7]. Regardless of the origin of these sinks, the interplay between defects and grain boundaries complicates our efforts to pinpoint the exact sites and chemistries of the entities present in the defective regions, thereby limiting our understanding of how specific defects mediate property changes. Here we show that the combination of advanced electron microscopy, spectroscopy and first-principles calculations can provide three-dimensional images of complex, multicomponent grain boundaries with both atomic resolution and chemical sensitivity. The high resolution of these techniques allows us to demonstrate that even for magnesium oxide, which has a simple rock-salt structure, grain boundaries can accommodate complex ordered defect superstructures that induce significant electron trapping in the bandgap of the oxide. These results offer insights into interactions between defects and grain boundaries in ceramics and demonstrate that atomic-scale analysis of complex multicomponent structures in materials is now becoming possible.

To understand clearly how the atomic-scale structure and composition of a buried interface affects the properties of a material requires experimentally resolving the sites and chemical identities of the atoms comprising the interface. In many cases, the presence of defects at grain boundaries is not known in advance. This poses a significant challenge to their identification because scattering and spectroscopy methods yield only an ensemble average of all defects[8], and scanning-probe microscopy often falls short of probing their chemical identities[9]. Advanced transmission electron microscopy (TEM) is in principle able to solve this problem[10,11], providing atomic-scale information on how defects are arranged in buried grain boundaries[12]. However, for complex grain boundaries, particularly those with multiple defects of uncertain species[13], precisely identifying each defect still poses a considerable challenge. Magnesium oxide (MgO), which is often considered a model oxide[14], presents such a case: it always contains trace amounts of impurities, some of which tend to segregate to grain boundaries[15,16]. These impurities are speculated to interact with native defects, dominating the structures of grain boundaries in MgO and hence determining many of its properties[17]. However, the three-dimensional structures of grain boundaries and how defects modify them are still unknown, even for the commonly occurring $\Sigma = 5$ grain boundaries[18] (Σ indicates the degree of geometrical coincidence at a grain boundary). These points remain a mystery owing to the intricacy of defect structures[19], the challenge of isolating specific grain boundaries[20] and the difficulty of interpreting individual atomic defects[21], despite their importance in understanding polycrystalline materials.

To address these issues, we fabricated a bicrystal (Fig. 1a, inset) with the bicrystallographic relations (310)[001]upper || (310)[00$\bar{1}$]lower; that is, each crystal is cut precisely along the (310) plane of MgO lattice, and the two revealed surfaces are joined with the [001] direction of the upper crystal parallel to the [00$\bar{1}$] direction of the lower crystal. Analysis of diffraction patterns confirms these orientations to within a small mis-fit tilt angle, and high-resolution TEM reveals a perfect join between the crystals at the atomic level (Supplementary Fig. 1). On closer inspection using aberration-corrected high-angle annular dark-field[1] (HAADF) scanning TEM (STEM), electron energy-loss

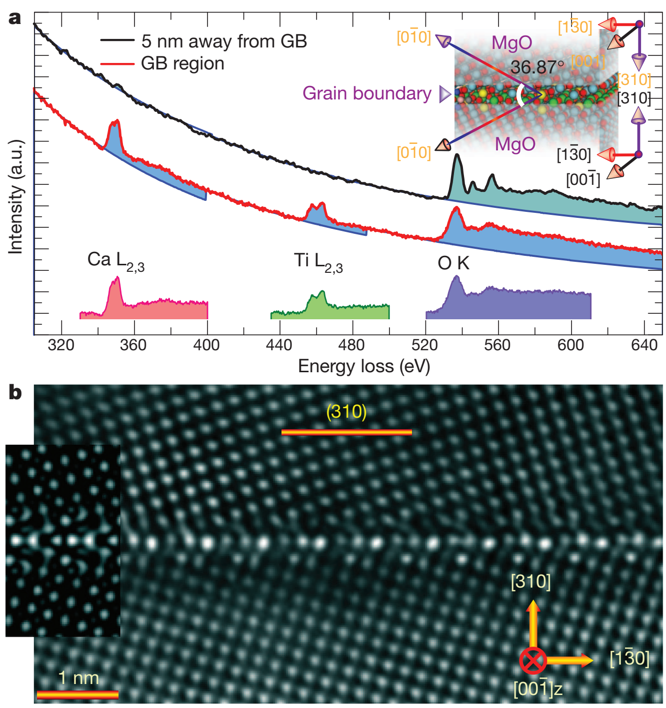


Figure 1 | Chemical and structural analysis of a $\Sigma = 5$, (310)[001] grain boundary. a, EELS spectrum from an energy-loss range containing the calcium $L_{2,3}$, titanium $L_{2,3}$ and oxygen K edges, showing the presence of calcium and titanium in the grain boundary (GB) region. The spectra with the background subtracted are shown at bottom. Inset, sketch of the MgO bicrystal. a.u., arbitrary units. **b**, High-resolution medium-voltage (400-kV) TEM image of the bicrystal viewed along the [001] direction. The grain boundary is atomically flat over extended regions up to several tens of nanometres in size. Comparing experimental images with simulated ones (inset) identifies the bright spots as sites of atomic columns.

[1]World Premier International Research Center, Advanced Institute for Materials Research, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan. [2]Department of Physics and Astronomy and London Centre for Nanotechnology, University College London, Gower Street, London WC1E 6BT, UK. [3]Institute of Engineering Innovation, The University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-8656, Japan. [4]Nanostructures Research Laboratory, Japan Fine Ceramics Center, 2-4-1 Mutsuno, Atsuta, Nagoya 456-8587, Japan.

spectroscopy[22] (EELS) and annular bright-field[23] (ABF) STEM, we obtain direct structural and element-selective imaging of the atoms comprising the grain boundaries, and demonstrate, with support of density functional theory calculations, that segregated impurities form an ordered defect superstructure that modifies electronic properties of MgO.

Figure 1a shows summed EELS spectra acquired from both the bulk and the region of the grain boundary, which provide definitive evidence of strong segregation of calcium and titanium into the otherwise impurity-free grain boundary. The presence of isovalent calcium is not surprising owing to the ion size mis-fit effect and because it is majority impurity in nominally pure MgO (ref. 17). However, we detected a comparable amount of titanium at the grain boundary, which had not been found previously (Supplementary Information). To extract exact segregation information, we present (Fig. 1b) a high-resolution TEM image of the grain boundary observed from the [001] direction. The boundary is atomically straight and the confined defects, which are represented as spots with variable image contrasts, are spatially periodic along the boundary's mirror plane. Further image simulations confirm that each bright spot corresponds to an atomic column[24] (Fig. 1b, inset).

This lateral atomic ordering at the grain boundary is unambiguously reflected in its corresponding HAADF STEM image (Fig. 2a), which shows spots with varying image contrasts inside periodic structural units (Fig. 2d). We also made a HAADF STEM image from the orthogonal $[1\bar{3}0]$ projection (Fig. 2b), which reveals a striped pattern (with periodicity a, the lattice constant of MgO) of higher contrast near the grain boundary. Because the intensity of an atomic column in the HAADF imaging mode is proportional to[25] $\sim Z^2$, where Z denotes atomic number, lighter spots away farther from the boundary represent normally pure MgO columns, whereas brighter pairs of spots in the boundary mirror plane represent MgO columns containing either calcium or titanium. Periodic spots with much lower image contrast in the mirror plane are also detected (Fig. 2a, arrows). This can be seen in an intensity line profile (Fig. 2c) between points I and II (Fig. 2a), which shows three peaks inside each structural unit. On closer inspection, the three spots are not composition equivalent, which poses a significant hurdle to determining individual defects at grain boundaries using HAADF imaging alone.

To clarify the chemical identity of the atoms at grain boundaries, we performed an atomic-resolution EELS analysis of the calcium $L_{2,3}$, titanium $L_{2,3}$ and oxygen K edges by focusing on an area containing the fundamental structural unit. A comparison of the HAADF (Fig. 2d, h) with calcium core-loss images (Fig. 2e, i) identifies the spots exactly on the boundary mirror plane as calcium-rich columns, which account for the brighter spots in both of the HAADF images. Titanium, on the other hand, resides dominantly at the spots nearest to the boundary mirror plane (Fig. 2f) and is periodically distributed in the $[1\bar{3}0]$ projection (Fig. 2j), which explains the striped contrast in the HAADF image (Fig. 2b). Conversely, the distribution of oxygen at the grain boundary is almost the same as that in the bulk (Fig. 2g, k), implying that calcium and titanium are chemically bonded to oxygen rather than forming metallic precipitates. Whereas calcium is isovalent to magnesium, previous studies on titanium-doped MgO concluded that titanium exists in a valence state of 3+ and is charge-compensated by negatively charged magnesium vacancies[26]. Although not directly detectable by HAADF imaging and EELS, the presence of such vacancies must be important in determining the structure and properties of the grain boundary.

To address this issue and provide a deeper understanding of the images, we conducted systematic density functional theory calculations, starting with a pure, $\Sigma = 5$, symmetric grain boundary (Fig. 3a). The stability of this boundary is enhanced by the introduction of calcium, which segregates to the mirror plane (Fig. 3b), inducing a translation of one grain with respect to the other by $a\sqrt{2/5}$ in the [310] direction. This calcium-doped grain boundary provides a good model for many of the features seen in the HAADF images, except that at the mirror plane each structural unit contains two spots of identical contrast rather than the observed three spots of varying contrast. To investigate whether titanium segregation and associated magnesium vacancies can be responsible for such a three-spot pattern, we examined preferred sites for titanium in all of its possible charge states (denoted $\mathrm{Ti}_{\mathrm{Mg}}^{\times}$, $\mathrm{Ti}_{\mathrm{Mg}}^{\bullet}$ and $\mathrm{Ti}_{\mathrm{Mg}}^{\bullet\bullet}$ in Kröger–Vink notation). In cases where the impurity is aliovalent titanium, we also considered the introduction of charge-compensating magnesium vacancies ($\mathrm{V}_{\mathrm{Mg}}^{\prime\prime}$) and determined their most stable arrangements within various atomic columns (Fig. 3b, labelled 1 to 11). For all three valence states, our energy calculations implied that segregation at column 1 is preferred, and in the cases where $\mathrm{V}_{\mathrm{Mg}}^{\prime\prime}$ is introduced we found that the vacancies still favour column 1, close to titanium (Supplementary Table 1), in remarkable consistence with the EELS mapping (Fig. 2f, j).

To determine which titanium valence state and defect configuration are stable, we further calculated the Gibbs free energy (γ) of the grain

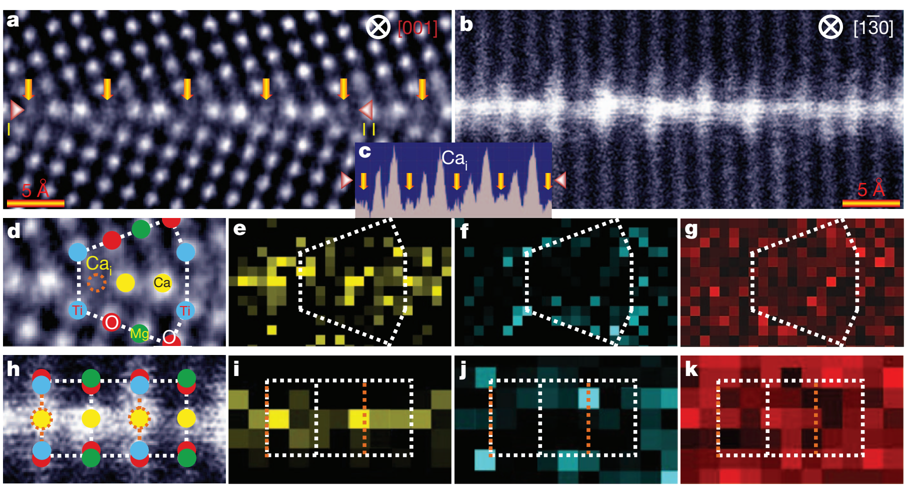


Figure 2 | Atomic-column imaging of the $\Sigma = 5$ grain boundary. **a**, **b**, Atomic-resolution HAADF images viewed from the [001] (**a**) and $[1\bar{3}0]$ (**b**) directions. The spots with low image contrast are indicated by arrows in **a**. We note that the distance between the (620) planes is as small as 0.66 Å, making them invisible in **b**. **c**, Line profile showing image intensity along the line I–II (**a**). The presence of entities giving rise to the spots with low image contrast is confirmed by clear peaks in the profile. **d**, **h**, Magnified HAADF images of the analysed region overlaid with the determined structural unit of the grain boundary observed from the [001] (**d**) and $[1\bar{3}0]$ (**h**) directions. In the structural unit, orange lines indicate planes containing titanium. **e**–**g**, **i**–**k**, Core-loss images of the calcium $L_{2,3}$, titanium $L_{2,3}$ and oxygen K edges viewed from the [001] (**e**–**g**) and $[1\bar{3}0]$ (**i**–**k**) directions. The core-loss images were made at the same places as the HAADF images in **d** and **h**.

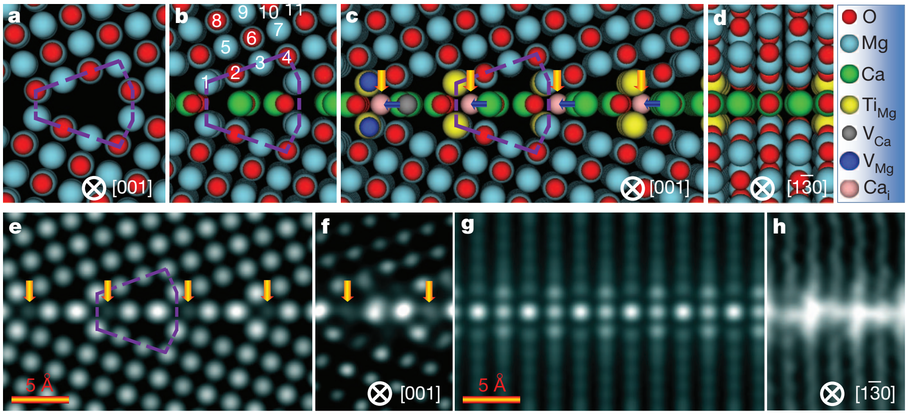


Figure 3 | Formation of an ordered defect superstructure at the grain boundary. **a**, Periodic supercell used to model the pure, $\Sigma = 5$ grain boundary. The boundary structural unit is indicated by the dashed polygon. **b**, The optimal model of the grain boundary with segregated calcium. The substitution and vacancy sites considered for titanium and, respectively, magnesium are numbered 1 to 11. **c**, **d**, The determined grain boundary viewed from the [001] (**c**) and [1$\bar{3}$0] (**d**) directions. The dots with low image contrast in Fig. 2a are recognized as calcium interstitials and are marked by arrows (yellow arrows show Ca$_i$ locations and blue arrows show from where the Ca atoms were displaced). **e**–**h**, Comparison showing a good match between simulated HAADF images obtained using the determined grain boundary and the corresponding low-pass-filtered images: [001] projection (**e**, **f**); [1$\bar{3}$0] projection (**g**, **h**).

boundaries as a function of the atomic chemical potentials of their constituents (μ_α), using the relation[27]

$$\gamma(\alpha,q) = \frac{1}{2A}\left[E(\text{defective},q) - \sum\nolimits_\alpha n_\alpha \mu_\alpha + qE_F\right]$$

where $E(\text{defective}, q)$ is the energy of the supercell with a defect in charge state q; n_α is the number of atoms of constituent α; E_F is the electron Fermi energy; and A is the interface area. Over a wide range of μ_O, the most stable grain boundary is that in which Ti^{3+} and V''_{Mg} both segregate to atomic column 1 (Fig. 4), which explains the absence of bright contrast in the HAADF images. The involvement of the 3+ valence state is corroborated by the measured electron energy-loss near-edge structure spectrum of the titanium $L_{2,3}$ edge, which is more consistent with the simulated spectrum of Ti^{3+} than with those of titanium's other valence states (Supplementary Fig. 2).

Geometrical investigation into the stable grain boundary revealed a structural transformation: calcium ions to the left of the pairs of bright spots in the structural units relax laterally towards interstitial sites (Fig. 4, inset). The driving force for such a displacement is the adjacent magnesium vacancies on both sides of the mirror plane, which free space into which calcium can relax to form a more stable site, substantially reducing the free energy of the grain boundary (Fig. 4). A similar effect is also found for the less stable grain boundary containing V''_{Mg}, indicating that the presence of interstitial calcium (Ca_i) can be viewed as an empirical test for the presence of V''_{Mg} at $\Sigma = 5$ grain boundaries. To provide further evidence for the existence of Ca_i, we performed ABF observations (Supplementary Fig. 3). These clearly resolve the Ca_i columns as dark dots at the interstitial sites.

In addition to yielding the spots with very low image contrast, the Ca_i atoms also leave behind calcium vacancies (V_{Ca}), thereby lowering the atomic density of their right-hand neighbours. This explains the difference in contrast between the two bright spots in each structural unit. We therefore find a new class of defect complexes emerging at grain boundaries: ordered defect superstructures comprising calcium and titanium impurities, magnesium vacancies, Ca_i and V_{Ca}. As a final confirmation, we simulated the HAADF images using the determined grain boundary (Fig. 3c, d) and compared them (Fig. 3e, g) with their low-pass-filtered counterparts (Fig. 3f, h). We found good agreement for both projections, providing further support for the conclusion that a grain boundary forms a complex multicomponent superstructure as a result of segregation and mutual interaction between defects of several species.

Atomistic imaging of an ordered defect superstructure at a grain boundary represents a step forward in our understanding of such boundaries because structural transformations of grain boundaries are

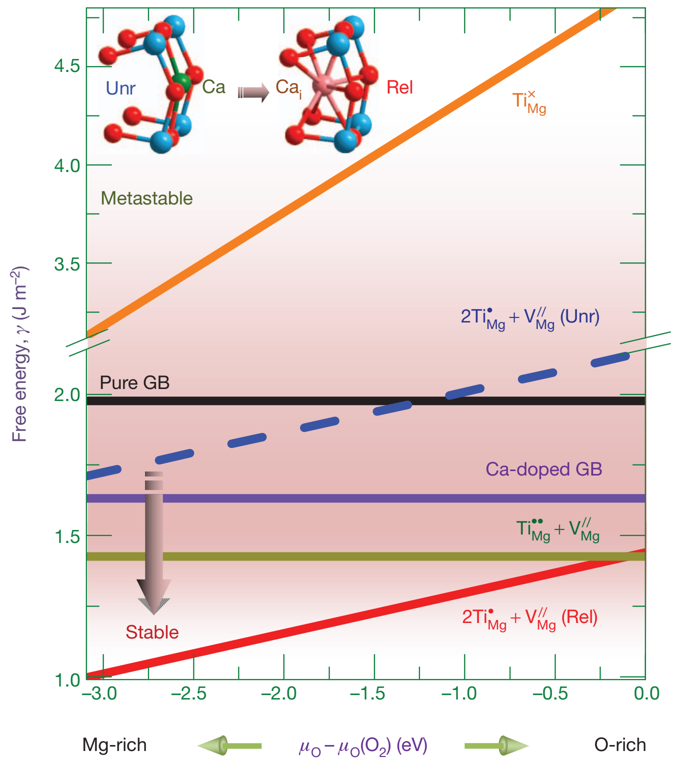


Figure 4 | Calculated free energy of grain boundary as a function of the chemical potential of oxygen (μ_O). We consider a pure MgO grain boundary, a calcium-doped grain boundary, and calcium-doped grain boundaries with, respectively, Ti^{2+} segregated to column 1 ($Ti^{\times}_{Mg}$), Ti^{3+} charge-compensated by V''_{Mg} at column 1 ($2Ti^{\bullet}_{Mg} + V''_{Mg}$ (Rel)) and Ti^{4+} charge compensated by V''_{Mg} at column 1 ($Ti^{\bullet\bullet}_{Mg} + V''_{Mg}$). The free energy of the $2Ti^{\bullet}_{Mg} + V''_{Mg}$ grain boundary before relaxation ($2Ti^{\bullet}_{Mg} + V''_{Mg}$ (Unr)) is given for comparison. The magnesium-rich and oxygen-rich ends of the chemical potential scale correspond to the cases in which MgO is in equilibrium with metallic magnesium and, respectively, O_2. Inset, structure transformation through optimization in the case of $2Ti^{\bullet}_{Mg} + V''_{Mg}$.

often considered to be induced by a high concentration of impurities[28]. Because the imaging of embedded defects is not yet routine, the prediction of properties of polycrystalline MgO relies largely on a widely held, simple model of a grain boundary with open defect cores[29,30] (Fig. 3a). The deviation of this model from ours raises the possibility that defects, even at low concentrations, can transform grain boundaries and affect their electronic properties, for instance the presence of deep electron traps (Supplementary Fig. 4). Our combined technique to resolve and identify defects in buried interfaces with unknown composition and structure should be applicable to other types of fundamental extended defect and a wide range of materials.

METHODS SUMMARY

The undoped, $\Sigma = 5$, (310)[001] symmetric tilt grain boundaries of MgO were fabricated using the bicrystal technique by joining two pristine MgO single crystals of high purity. Specimens for TEM and STEM observations were prepared by cutting, grinding, dimpling and argon ion-beam thinning. The HAADF images were observed with a 200-kV STEM (JEM-2100F, JEOL) equipped with an aberration corrector (CEOS Gmbh), which offers an unprecedented opportunity to probe structures with sub-ångström resolution. The ABF images were made with a 6–25-mrad detector, and the EELS spectra were recorded using a Gatan ENFINA system equipped on STEM with an energy resolution (full-width of half-maximum) of $\sim$0.9–0.95 eV. Image simulations were performed using the WinHREM program (HREM Research Inc.), which is based on the multislice method[24] and takes into account the absorption of thermal diffuse scattering for each element. Density functional theory calculations were carried out using the VASP code. We applied the projector augmented-wave method with $4 \times 4 \times 1$ k-point grids and a cut-off energy of 400 eV, and the GGA + U method with $U = 3.0$ and $J = 1.0$ eV for Ti d states. The grain boundary was modelled both as a periodic supercell of size 8.47 Å $\times$ 13.24 Å $\times$ 31.51 Å and as a double supercell to examine size effects as well as conduct image simulations (Supplementary Information).

Received 30 June; accepted 21 September 2011.

1. Nellist, P. D. *et al.* Direct sub-angstrom imaging of a crystal lattice. *Science* **305,** 1741 (2004).
2. Buban, J. P. *et al.* Grain boundary strengthening in alumina by rare earth impurities. *Science* **311,** 212–215 (2006).
3. Kingery, W. D. Plausible concepts necessary and sufficient for interpretation of ceramic grain-boundary phenomena: I, grain-boundary characteristics, structure, and electrostatic potential. *J. Am. Ceram. Soc.* **57,** 1–8 (1974).
4. Bai, X. M. *et al.* Efficient annealing of radiation damage near grain boundaries via interstitial emission. *Science* **327,** 1631–1634 (2010).
5. Jia, C. L. & Urban, K. Atomic-resolution measurement of oxygen concentration in oxide materials. *Science* **303,** 2001–2004 (2004).
6. Lartigue-Korinek, S., Bouchet, D., Bleloch, A. & Colliex, C. HAADF study of the relationship between intergranular defect structure and yttrium segregation in an alumina grain boundary. *Acta Mater.* **59,** 3519–3527 (2011).
7. Maiti, A. *et al.* Dopant segregation at semiconductor grain boundaries through cooperative chemical rebonding. *Phys. Rev. Lett.* **77,** 1306–1309 (1996).
8. Kaiser, U. *et al.* Direct observation of defect-mediated cluster nucleation. *Nature Mater.* **1,** 102–105 (2002).
9. Barth, C. *et al.* Recent trends in surface characterization and chemistry with high-resolution scanning force methods. *Adv. Mater.* **23,** 477–501 (2011).
10. Muller, D. A. Structure and bonding at the atomic scale by scanning transmission electron microscopy. *Nature Mater.* **8,** 263–270 (2009).
11. Kimoto, K. *et al.* Element-selective imaging of atomic columns in a crystal using STEM and EELS. *Nature* **450,** 702–704 (2007).
12. Huang, P. Y. *et al.* Grains and grain boundaries in single-layer graphene atomic patchwork quilts. *Nature* **469,** 389–392 (2011).
13. Klie, R. F. *et al.* Enhanced current transport at grain boundaries in high-T_c superconductors. *Nature* **435,** 475–478 (2005).
14. McKenna, K. P. & Shluger, A. L. First-principles calculations of defects near a grain boundary in MgO. *Phys. Rev. B* **79,** 224116 (2009).
15. Chiang, Y. M., Henriksen, A. F. & Kingery, W. D. Characterization of grain-boundary segregation in MgO. *J. Am. Ceram. Soc.* **64,** 385–389 (1981).
16. Browning, N. D. *et al.* Investigating the structure-property relationships at grain boundaries in MgO using bond-valence pair potentials and multiple scattering analysis. *J. Am. Ceram. Soc.* **82,** 366–372 (1999).
17. Yan, Y. *et al.* Impurity-induced structural transformation of a MgO grain boundary. *Phys. Rev. Lett.* **81,** 3675–3678 (1998).
18. Duffy, D. M. Grain boundaries in ionic crystals. *J. Phys. C* **19,** 4393–4412 (1986).
19. Yamakov, V. *et al.* Dislocation processes in the deformation of nanocrystalline aluminium by molecular-dynamics simulation. *Nature Mater.* **1,** 45–49 (2002).
20. Kizuka, T., Iijima, M. & Tanaka, N. Atomic process of electron-irradiation-induced grain-boundary migration in a MgO tilt boundary. *Philos. Mag. A* **77,** 413–422 (1998).
21. Ortalan, V. *et al.* Direct imaging of single metal atoms and clusters in the pores of dealuminated HY zeolite. *Nature Nanotechnol.* **5,** 506–510 (2010).
22. Colliex, C. Elementary resolution. *Nature* **450,** 622–623 (2007).
23. Findlay, S. *et al.* Robust atomic resolution imaging of light elements using scanning transmission electron microscopy. *Appl. Phys. Lett.* **95,** 191913 (2009).
24. Ishizuka, K. & Uyeda, N. A new theoretical and practical approach to the multislice methods. *Acta Crystallogr. A* **33,** 740–749 (1977).
25. Pennycook, S. J. & Boatner, L. A. Chemically sensitive structure-imaging with a scanning transmission electron microscope. *Nature* **336,** 565–567 (1988).
26. Davies, J. J. & Wertz, J. E. The EPR spectrum of trivalent titanium in orthorhombic symmetry in MgO. *J. Phys. Chem. Solids* **31,** 2489–2494 (1970).
27. Tanaka, I. *et al.* Identification of ultradilute dopants in ceramics. *Nature Mater.* **2,** 541–545 (2003).
28. Vitek, V. & Wang, G. J. Segregation and grain boundary structure. *Surf. Sci.* **144,** 110–123 (1984).
29. McKenna, K. P. & Shluger, A. L. Electron-trapping polycrystalline materials with negative electron affinity. *Nature Mater.* **7,** 859–862 (2008).
30. Harris, D. J., Watson, G. W. & Parker, S. C. Atomistic simulation of the effect of temperature and pressure on the [001] symmetric tilt grain boundaries of MgO. *Phil. Mag. A* **74,** 407–418 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Area "Atomic Scale Modification (474)" from MEXT, Japan. We thank T. Mizoguchi for performing electron energy-loss near-edge structure simulations and for discussions, and T. Saito and W. Zeng for experimental assistance. Z.W. acknowledges support by a Grant-in-Aid for Young Scientists (B) (grant no. 22760500) and from IZUMI Science Foundation. M.S. is grateful for a Grant-in-Aid for Scientific Research (C) (grant no. 23560817) and to MURATA Science Foundation for financial support. K.P.M. acknowledges support by a Grant-in-Aid for Young Scientists (B) (grant no. 22740192). S.T. thanks supports from the Nippon Sheet Glass Foundation. Calculations were conducted at ISSP, University of Tokyo.

Author Contributions Z.W. prepared specimens, carried out calculations and wrote the manuscript. M.S. made images and conducted image simulation and processing. K.P.M. and A.L.S. helped with the calculations and discussed the results. L.G. and S.T. helped with the experiments. Y.I. discussed the results and directed the entire study. All the authors read and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to Z.W. (zcwang@wpi-aimr.tohoku.ac.jp).

LETTER

doi:10.1038/nature10588

Observed increase in local cooling effect of deforestation at higher latitudes

Xuhui Lee[1], Michael L. Goulden[2], David Y. Hollinger[3], Alan Barr[4], T. Andrew Black[5], Gil Bohrer[6], Rosvel Bracho[7], Bert Drake[8], Allen Goldstein[9], Lianhong Gu[10], Gabriel Katul[11], Thomas Kolb[12], Beverly E. Law[13], Hank Margolis[14], Tilden Meyers[15], Russell Monson[16], William Munger[17], Ram Oren[11], Kyaw Tha Paw U[18], Andrew D. Richardson[19], Hans Peter Schmid[20], Ralf Staebler[21], Steven Wofsy[17] & Lei Zhao[1]

Deforestation in mid- to high latitudes is hypothesized to have the potential to cool the Earth's surface by altering biophysical processes[1–3]. In climate models of continental-scale land clearing, the cooling is triggered by increases in surface albedo and is reinforced by a land albedo–sea ice feedback[4,5]. This feedback is crucial in the model predictions; without it other biophysical processes may overwhelm the albedo effect to generate warming instead[5]. Ongoing land-use activities, such as land management for climate mitigation, are occurring at local scales (hectares) presumably too small to generate the feedback, and it is not known whether the intrinsic biophysical mechanism on its own can change the surface temperature in a consistent manner[6,7]. Nor has the effect of deforestation on climate been demonstrated over large areas from direct observations. Here we show that surface air temperature is lower in open land than in nearby forested land. The effect is 0.85 ± 0.44 K (mean ± one standard deviation) northwards of 45° N and 0.21 ± 0.53 K southwards. Below 35° N there is weak evidence that deforestation leads to warming. Results are based on comparisons of temperature at forested eddy covariance towers in the USA and Canada and, as a proxy for small areas of cleared land, nearby surface weather stations. Night-time temperature changes unrelated to changes in surface albedo are an important contributor to the overall cooling effect. The observed latitudinal dependence is consistent with theoretical expectation of changes in energy loss from convection and radiation across latitudes in both the daytime and night-time phase of the diurnal cycle, the latter of which remains uncertain in climate models[8].

The latitudinal gradient of land-use impact is evident in the comparison of the surface air temperature recorded at FLUXNET (www.fluxnet.ornl.gov) forest towers[9] (Supplementary Table 1 and Supplementary Fig. 1) and surface weather stations in North America (Fig. 1a). Here we use the surface stations as proxies for cleared land. In accordance with the requirement of the World Meteorological Organization, these stations are located in open grassy fields that have biophysical characteristics similar to those of open land, such as being covered by snow in northern latitudes in the winter[10]. Latitude accounts for 31% of the variations in the temperature difference ΔT between the forest sites and the adjacent open lands (number of site pairs $n = 37$). The rate of change in ΔT with latitude is -0.070 ± 0.010 K per degree (mean ± one standard error, s.e., $P < 0.005$). At these sites, the annual net all-wave radiation R_n

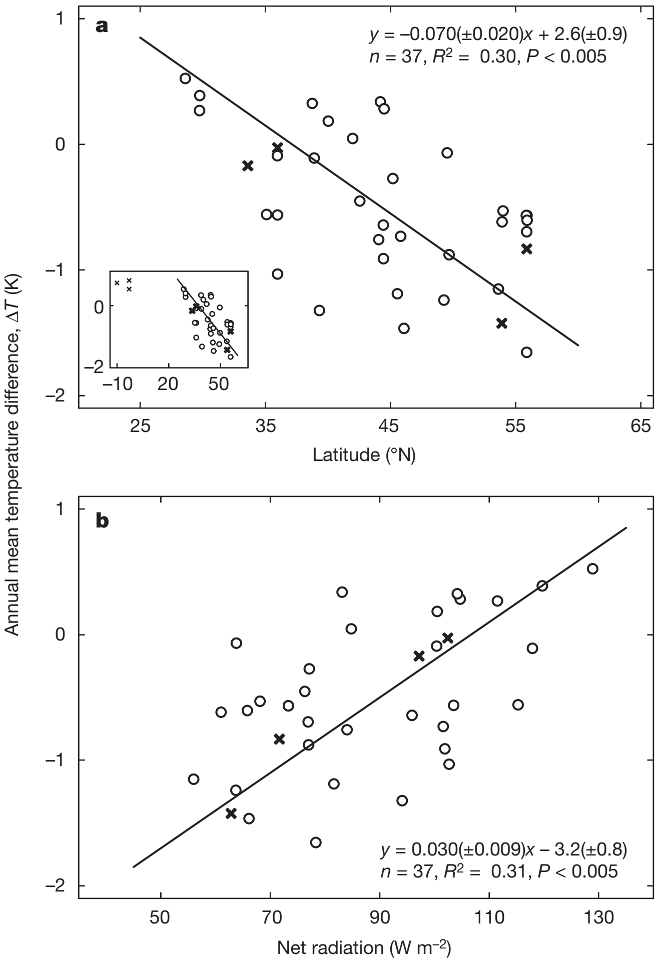


Figure 1 | Annual mean difference (open land minus forest) in surface air temperature. a, Correlation with latitude. **b**, Correlation with surface net radiation. The inset to **a** has the same axes as the main panel but also shows tropical FLUXNET site data. Parameter bounds in the linear regression are for the 95% confidence interval. Circles indicate weather station/forest site pairs and crosses indicate FLUXNET site clusters.

[1]School of Forestry and Environmental Studies, Yale University, New Haven, Connecticut 06511, USA. [2]Department of Earth System Science, University of California, Irvine, California 92697, USA. [3]USDA Forest Service, Northern Research Station, Durham, New Hampshire 03824, USA. [4]Climate Research Division, Environment Canada, Saskatoon, S7N 3H5, Canada. [5]Faculty of Land and Food Systems, University of British Columbia, Vancouver, V6T 1Z4, Canada. [6]Department of Civil and Environmental Engineering and Geodetic Science, Ohio State University, Columbus, Ohio 43210, USA. [7]School of Forest Resources and Conservation, University of Florida, Gainesville, Florida 32611, USA. [8]Smithsonian Environmental Research Center, Edgewater, Maryland 21037, USA. [9]Department of Environmental Science, Policy and Management, University of California, Berkeley, California 94720, USA. [10]Environmental Science Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. [11]Nicholas School of the Environment and Earth Science, Duke University, Durham, North Carolina 27708, USA. [12]School of Forestry, Northern Arizona University, Flagstaff, Arizona 86011, USA. [13]College of Forestry, Oregon State University, Corvallis, Oregon 97331, USA. [14]Centre d'Étude de la Forêt, Faculté de Foresterie, de Géographie et de Géomatique, Université Laval, Québec City, Québec, G1V 0A6, Canada. [15]NOAA/ARL/ATDD, Oak Ridge, Tennessee 37830, USA. [16]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309, USA. [17]School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA. [18]Department of Land, Air and Water Resources, University of California, Davis, California 95616, USA. [19]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [20]Institute of Meteorology and Climate Research, Karlsruhe Institute for Technology, 82467 Garmisch-Partenkirchen, Germany. [21]Processes Research Section, Environment Canada, Toronto, Ontario, M3H 5T4, Canada.

decreases linearly with latitude. If R_n is used as the independent variable, the correlation becomes positive (Fig. 1b, linear correlation 0.55, $P < 0.005$). Combining the site-pair observations with the limited tropical FLUXNET data suggests that the latitudinal dependence may level off in zones south of the paired analysis (inset to Fig. 1a).

For the site pairs north of 45° N (Fig. 2a), the mean annual ΔT is -0.85 ± 0.44 K (mean ± 1 standard deviation, s.d.), a result in agreement with, but weaker than, those of climate model simulations of large-scale land-use changes in the boreal zone[11,12]. If we approximate the net shortwave radiation change at these site pairs by the boreal FLUXNET site cluster data (Supplementary Table 2), we arrive at an apparent local climate sensitivity of about 0.027 K W^{-1} m^2. The mean monthly ΔT does not seem to depend on season (Fig. 2a); the modelled maximum temperature change from March to May[12,13] is not discernible in our data, suggesting some strong compensating signals in the real atmosphere.

For the site pairs south of 45° N (Fig. 2b), the mean annual ΔT is -0.21 ± 0.53 K, giving an apparent sensitivity of about 0.012 K W^{-1} m^2. There appears to exist a weak seasonality, with the open sites cooler than the forests (−0.52 K) in January and slightly warmer (0.08 K) in June. For comparison, the cooling signal associated with historical land clearing since the 1700s, which has occurred primarily in mid-latitudes[13], is 0.5–1.0 K.

The latitudinal dependence can be understood by examining the intrinsic biophysical mechanism. Forests have lower surface albedo than shrubs, grasses and pastures[6,7,14]. Deforestation decreases the net shortwave absorption by an amount ΔS that depends in part on climate regimes. Local surface temperature would fall in response to the decreased surface radiation loading associated with deforestation if radiation were the only energy transfer process involved. Similar to the global analysis[15,16], the surface temperature change would be

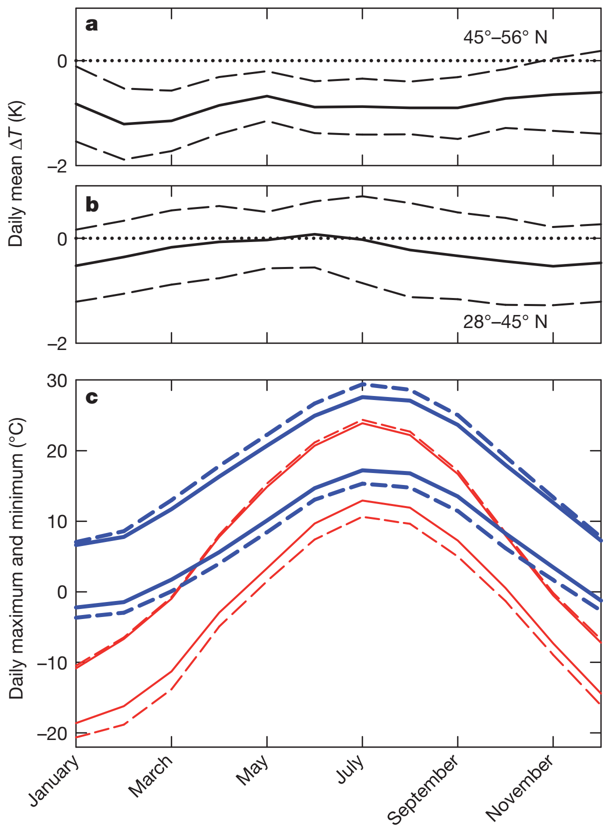


Figure 2 | Seasonal and diurnal patterns of surface air temperature. **a** and **b** show the mean temperature difference ±1 s.d. for the site pairs north and south of 45° N. **c**, Mean daily maximum and minimum temperatures for the forests (solid lines) and the surface stations (dotted lines) for 28–45° N (blue lines) and 45–56° N (red lines).

$\Delta T_s = \lambda_0 \Delta S$, where $\Delta S < 0$ and the temperature sensitivity resulting from the longwave radiation feedback $\lambda_0 = 1/(4\sigma T_s^3) \approx 0.2$ K W^{-1} m^2. (T_s is surface temperature and σ is the Stephan–Boltzmann constant.) The actual temperature change also depends on energy redistribution through convection and evapotranspiration. Owing to their larger aerodynamic roughness, forests dissipate sensible heat more efficiently to the atmospheric boundary layer than do open landscapes[6]. In humid climates, they also remove from the surface more latent heat[14], which is released above the atmospheric boundary layer by cloud condensation.

The intrinsic biophysical mechanism can be expressed as a temperature change in response to changes in these energy exchange processes:

$$\Delta T_s \approx \lambda_0 \Delta S/(1+f) + (-\lambda_0) R_n \Delta f/(1+f)^2 \qquad (1)$$

where f (>0) is an energy redistribution factor. Equation (1) reveals a number of useful properties of the biophysical effect. The first term on the right (radiative forcing term) results from albedo changes but is always damped by energy redistribution. The second term (energy redistribution) has two additive components contributed by changes in Bowen ratio and in surface roughness and over the diurnal cycle is usually positive when forests are converted to open land. Because these terms have opposite signs, the local climate sensitivity with respect to ΔS cannot exceed the upper limit of λ_0 and can even be negative[6]. Equation (1) calls attention to a previously unrecognized role of R_n, which is to amplify the effects of roughness and Bowen ratio changes in low latitudes and reduce these effects in high latitudes. Because with increasing latitude ΔS becomes more negative[2] and R_n decreases, equation (1) suggests that ΔT_s should be negatively correlated with latitude (Fig. 1a).

A conceptual analysis using equation (1) suggests that the relative contribution of the different biophysical forcings of deforestation to ΔT_s should depend on the climate zone (Fig. 3 and Supplementary Table 2). In the boreal zone, open land from timber harvest generates stronger radiative cooling and roughness warming (Fig. 3a) than fire (Fig. 3b). This is because the standing dead trees in the recently burnt site[17] partially mask winter snow cover and enhance turbulent mixing year round. The temperate site cluster (Fig. 3c) displays a weaker radiative cooling but stronger roughness warming than the boreal clusters. Over the broad parameter space shown in Fig. 3, the partitioning of the net temperature change bears remarkable resemblance to the results of a global-scale deforestation experiment in a climate model, but only after the sea-ice feedback has been included in the model[5].

Surprisingly, a diurnal asymmetry exists in the biophysical effect (Fig. 2c). Diurnal temperature range (DTR, the difference between the daily maximum and minimum temperature) is an important measure of surface climate variability[18,19] and we find that DTR is reduced with forest cover. At night, ΔS vanishes, surface evapotranspiration is generally negligible, and surface roughness is the main biophysical factor affecting the daily minimum temperature changes (Supplementary equation (14), noting $R_n < 0$). Even though the lower roughness contributes to a warming of the daily mean temperature (Fig. 3), at night open land cools more than forests in both the northern and the southern latitudes. We hypothesize that forests are warmer at night because in stable stratification the presence of trees causes turbulence, bringing heat from aloft to the surface. The mechanism underlying the daily maximum temperature changes is more complex. In the daytime, suppressed mixing due to a smaller surface roughness causes the surface temperature of the open land to rise faster than that of the forest. At the sites north of 45° N, this roughness effect is, however, nearly offset by cooling associated with albedo and Bowen ratio changes, resulting in almost identical daily maximum temperatures between the paired sites. The diurnal asymmetry emphasizes the importance of both daytime and night-time observations for obtaining an accurate assessment of the land-use effect.

Forests represent one of the most extensive land-use types, occupying about 30% of the terrestrial surface[1]. Our paired analysis

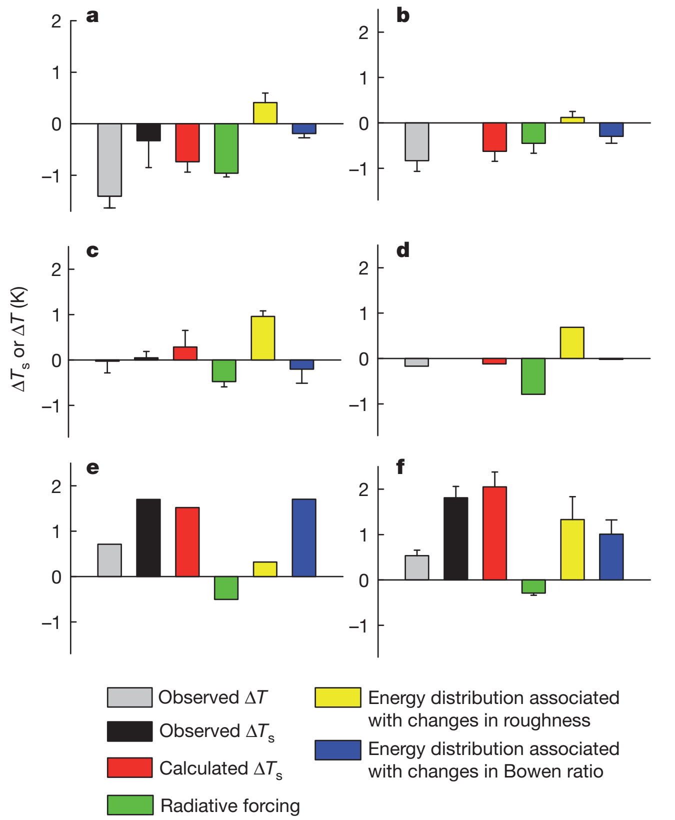


Figure 3 | Partition of the biophysical effect at six FLUXNET site clusters in four different climate zones. a, Boreal: harvested site versus jack pine forests[25]. **b**, Boreal: burnt site versus black spruce forests[17]. **c**, Temperate: grassland versus pine and oak/hickory forests[26]. **d**, Semi-arid: open shrub land versus pinyon juniper[27]. **e**, Tropical: pasture versus rainforest[14]. **f**, Tropical: farmland versus rainforest[28,29]. Temperatures are 24-h means. Error bars are given as 1 s.d. for the clusters with multiple site-year observations. No surface temperature measurements are available for **b** and **d**. For comparison, observed changes in surface air temperature (ΔT) are also shown.

indicates that biases exist in the climate station observations if they are used to represent the climate state of the forested land. These biases are manifested in several ways. First, the station network overestimates the north–south surface temperature gradient by 0.070 ± 0.010 K per °N (mean ± 1 s.e.) for this land type (Fig. 1a). Second, DTR is a variable sensitive to the biophysical properties of the surface above which the observation is made. The station DTR is biased high by an average of 2.8 ± 2.0 K (and up to 8 K in some locations) in comparison to that at

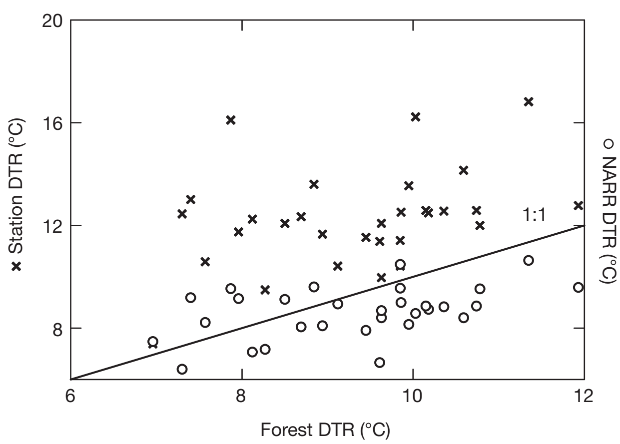


Figure 4 | Comparison of the DTR for the forests, the surface stations and the NARR model result.

the FLUXNET forests (Fig. 4). Third, these station biases can adversely affect model-data comparison. The models in question include both climate models and those used for atmospheric data assimilation. The modelled screen-height temperature is not compatible with the station observations because the grids prescribed as forest vegetation in the model domain have different biophysical properties from the surface of a weather station. For example, the DTR modelled by NARR (North American Regional Reanalysis)[20] is in much better agreement with the FLUXNET observations than with the surface station data (Fig. 4). Similar to the observed pattern (Fig. 1a), the station temperature becomes progressively lower with increasing latitude than the NARR-predicted screen-height (2.0 m above the vegetation surface) temperature (Supplementary Fig. 2). The model-data incompatibility may be one reason for why the DTR trends simulated by climate models do not agree with the observed trends[18,21], although firm evidence for this will need longer FLUXNET temperature records. It also provides additional evidence showing that the intrinsic biophysical mechanism can alter surface air temperatures in a predictive manner without influences originating from outside the atmospheric boundary layer.

The purpose of this analysis is to quantify the biophysical effect arising from spatial patterns of land use in the present climate. We assume that the adjacent land types share the same background state defined by the incoming solar radiation, the incoming longwave radiation and air temperature at the blending height[22] above the ground. Substitution of the spatial variations for time variations of land use must also consider that this background state may be changing due to large-scale variations in radiative forcing and climate feedbacks[23,24]. Conceptually, the intrinsic biophysical processes can be regarded as a local perturbation superimposed on the changing background. We postulate that at scales of ongoing land-use activities, the perturbation signals are much larger than the background changes. For example, clearing of a million hectares of forest (the size of a typical climate model grid) would reduce the global radiative forcing associated with albedo changes by 3×10^{-4} W m^{-2} (ref. 2), which is too weak to cause observable changes in surface temperature.

METHODS SUMMARY

FLUXNET and station data. Data obtained at 33 FLUXNET forest sites in the USA and Canada are used in this analysis. These sites have a minimum of three years of temperature and net radiation data. The surface weather station closest to every forest site was chosen for the paired analysis. The site pairs have a mean elevation difference of 59 m, a linear distance of 28 km and a latitudinal distance of 0.2 km. The height of the temperature measurement in the FLUXNET network varies from 2 to 15 m above the canopy. Correcting the measurement to the standard screen height (2.0 m above the vegetation) would change the annual mean temperature by no more than 0.1 K.

NARR data. NARR uses the NCEP (National Centers for Environmental Protection) Eta model and numerous data sources to produce outputs at a grid spacing of 32 km. Surface station observations of the screen-height temperature are not used to constrain the modelled fields. Each forest site is matched up with the closest NARR grid. At these grids, calculations of the NARR screen-height temperature are forced with a surface boundary with biophysical properties of forested landscapes.

Model of biophysical processes. Equation (1) is derived from a linearized version of the surface energy balance equation. It was assumed that in the vicinity of one another, a forest area and a piece of open land receive the same amounts of incoming shortwave and longwave radiation and that air is sufficiently mixed at the blending height. As a result, any difference in the surface temperature is caused by the intrinsic biophysical mechanism or changes in albedo, surface roughness and Bowen ratio. Evaluation of equation (1) for the six FLUXNET site clusters (Fig. 3) was done separately for the daytime and night-time periods to avoid nonlinear parameter interactions through the diurnal cycle.

Received 5 June; accepted 22 September 2011.

1. Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* **320,** 1444–1449 (2008).
2. Betts, R. A. Offset of the potential carbon sink from boreal forestation by decreases in surface albedo. *Nature* **408,** 187–190 (2000).

3. Bala, G. *et al.* Combined climate and carbon cycle effects of large-scale deforestation. *Proc. Natl Acad. Sci. USA* **104,** 6550–6555 (2007).
4. Bonan, G. B. *et al.* Effects of boreal forest vegetation on global climate. *Nature* **359,** 716–718 (1992).
5. Davin, E. L. & De Noblet-Ducoudré, N. Climatic impact of global-scale deforestation: radiative versus nonradiative processes. *J. Clim.* **23,** 97–112 (2010).
6. Rotenberg, E. & Yakir, D. Contribution of semi-arid forests to the climate system. *Science* **327,** 451–454 (2010).
7. Juang, J.-Y. *et al.* Separating the effects of albedo from eco-physiological changes on surface temperature along a successional chronosequence in the southeastern United States. *Geophys. Res. Lett.* **34,** L21408 (2007).
8. Pielke, R. A. *et al.* Unresolved issues with the assessment of multidecadal global land surface temperature trends. *J. Geophys. Res.* **112,** D24S08 (2007).
9. Baldocchi, D. *et al.* FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor and energy flux densities. *Bull. Am. Meteorol. Soc.* **82,** 2415–2434 (2001).
10. Betts, A. K. & Ball, J. H. Albedo over the boreal forest. *J. Geophys. Res.* **102, D24,** 28901–28909 (1997).
11. Douville, H. & Royer, J.-F. Influence of the temperate and boreal forests on the Northern Hemisphere climate in the Meteo-France climate model. *Clim. Dyn.* **13,** 57–74 (1997).
12. Snyder, P. K. *et al.* Evaluating the influence of different vegetation biomes on the global climate. *Clim. Dyn.* **23,** 279–302 (2004).
13. Betts, R. A. *et al.* Biogeophysical effects of land use on climate: model simulations of radiative forcing and large-scale temperature change. *Agric. For. Meteorol.* **142,** 216–233 (2007).
14. von Randow, C. *et al.* Comparative measurements and seasonal variations in energy and carbon exchange over forest and pasture in south west Amazonia. *Theor. Appl. Climatol.* **78,** 5–26 (2004).
15. Pielke, R. A. & Avissar, R. Influence of landscape structure on local and regional climate. *Landscape Ecol.* **4,** 133–155 (1990).
16. Hansen, J. *et al.* in *Climate Processes and Climate Sensitivity* (eds Hansen, J. E. & Takahashi, T.) 130–163 (American Geophysical Union, 1984).
17. Goulden, M. L. *et al.* An eddy covariance mesonet to measure the effect of forest age on land-atmosphere exchange. *Glob. Change Biol.* **12,** 2146–2162 (2006).
18. Wild, M. *et al.* Impact of global dimming and brightening on global warming. *Geophys. Res. Lett.* **34,** L04702 (2007).
19. Easterling, D. R. *et al.* Maximum and minimum temperature trends for the globe. *Science* **277,** 364–367 (1997).
20. Mesinger, F. *et al.* North American regional reanalysis. *Bull. Am. Meteorol. Soc.* **87,** 343–360 (2006).
21. Zhou, L. *et al.* Spatiotemporal patterns of changes in maximum and minimum temperatures in multi-model simulations. *Geophys. Res. Lett.* **36,** L02702 (2009).
22. Mahrt, L. Surface heterogeneity and vertical structure of the boundary layer. *Boundary-Layer Meteorol.* **96,** 33–62 (2000).
23. Swann, A. L. *et al.* Changes in Arctic vegetation amplify high-latitude warming through the greenhouse effect. *Proc. Natl Acad. Sci. USA* **107,** 1295–1300 (2010).
24. Davin, E. L. *et al.* Impact of land cover change on surface climate: relevance of radiative forcing concept. *Geophys. Res. Lett.* **34,** L13702 (2007).
25. Zha, T. *et al.* Carbon sequestration in boreal jack pine stands following harvesting. *Glob. Change Biol.* **15,** 1475–1487 (2009).
26. Stoy, P. C. *et al.* Separating the effects of climate and vegetation on evapotranspiration along a successional chronosequence in the southeastern US. *Glob. Change Biol.* **12,** 2115–2135 (2006).
27. Anderson, R. G. & Goulden, M. L. Relationships between climate, vegetation, and energy exchange across a montane gradient. *J. Geophys. Res.* **116,** G01026 (2011).
28. Goulden, M. L. *et al.* Diel and seasonal patterns of tropical forest CO_2 exchange. *Ecol. Appl.* **14,** 42–54 (2004).
29. Sakai, R. *et al.* Land-use change effects on local energy, water, and carbon balances in an Amazonian agricultural field. *Glob. Change Biol.* **10,** 895–907 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The data collection and analysis were supported in part by grants from the US Department of Energy and by a Yale University Climate and Energy Institute grant. We thank D. Fitzjarrald and R. Sakai for providing the data for the KM77 tropical site and C. von Randow for providing the friction velocity data for FLUXNET cluster e.

Author Contributions X.L. developed the energy balance model, carried out the analysis and wrote the manuscript, M.L.G. and D.Y.H. contributed ideas to data analysis, M.L.G., D.Y.H., T.A.B., G.B., L.G., G.K., T.K., B.E.L., H.M., T.M., W.M., R.O., A.D.R., R.S. and S.W. contributed ideas to manuscript development, M.L.G., D.Y.H., A.B., T.A.B., G.B., R.B., B.D., A.G., L.G., G.K., T.K., B.E.L., X.L., H.M., T.M., R.M., W.M., R.O., K.T.P.U, A.D.R., H.P.S., R.S. and S.W. contributed data, and L.Z. performed the NARR data analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to X.L. (xuhui.lee@yale.edu).

doi:10.1038/nature10566

East Antarctic rifting triggers uplift of the Gamburtsev Mountains

Fausto Ferraccioli[1], Carol A. Finn[2], Tom A. Jordan[1], Robin E. Bell[3], Lester M. Anderson[1] & Detlef Damaske[4]

The Gamburtsev Subglacial Mountains are the least understood tectonic feature on Earth, because they are completely hidden beneath the East Antarctic Ice Sheet. Their high elevation and youthful Alpine topography, combined with their location on the East Antarctic craton, creates a paradox that has puzzled researchers since the mountains were discovered in 1958[1]. The preservation of Alpine topography in the Gamburtsevs[2] may reflect extremely low long-term erosion rates beneath the ice sheet[3], but the mountains' origin remains problematic. Here we present the first comprehensive view of the crustal architecture and uplift mechanisms for the Gamburtsevs, derived from radar, gravity and magnetic data. The geophysical data define a 2,500-km-long rift system in East Antarctica surrounding the Gamburtsevs, and a thick crustal root[4] beneath the range. We propose that the root formed during the Proterozoic assembly of interior East Antarctica (possibly about 1 Gyr ago), was preserved as in some old orogens[5,6] and was rejuvenated during much later Permian (roughly 250 Myr ago) and Cretaceous (roughly 100 Myr ago) rifting. Much like East Africa[7], the interior of East Antarctica is a mosaic of Precambrian provinces affected by rifting processes. Our models show that the combination of rift-flank uplift, root buoyancy and the isostatic response to fluvial and glacial erosion explains the high elevation and relief of the Gamburtsevs. The evolution of the Gamburtsevs demonstrates that rifting and preserved orogenic roots can produce broad regions of high topography in continental interiors without significantly modifying the underlying Precambrian lithosphere.

Although the Gamburtsevs have been identified as a site of early Antarctic ice-sheet growth[2], their age and origin remain a matter of considerable speculation. A Precambrian basement for the Gamburtsevs has been inferred from outcrops to the north[8], detrital zircon data[9] and thick, high-velocity lithosphere[10] (Fig. 1). Such lithosphere is uniquely associated with Precambrian cratons that have not been significantly deformed by later subduction and collision. The high elevation and relief of the Gamburtsevs is, however, highly anomalous, given their location in the interior of a craton. Proposed models for the origin of the Gamburtsevs include: Neoproterozoic–early Cambrian orogenic events associated with Gondwana assembly[11]; far-field compression linked with Pangaea formation during the late Carboniferous–early Permian periods[12]; and uplift over a Cenozoic mantle plume[13]. The lack of modern geophysical exploration and drilling in central East Antarctica has hindered efforts to validate these models or propose new ones.

Understanding the origin of the Gamburtsevs was a primary goal of the seven-nation International Polar Year Antarctica's Gamburtsev Provice (AGAP) project that explored central East Antarctica using two Twin Otter aircraft equipped with ice-penetrating radars, laser ranging systems, gravity meters and magnetometers. More than

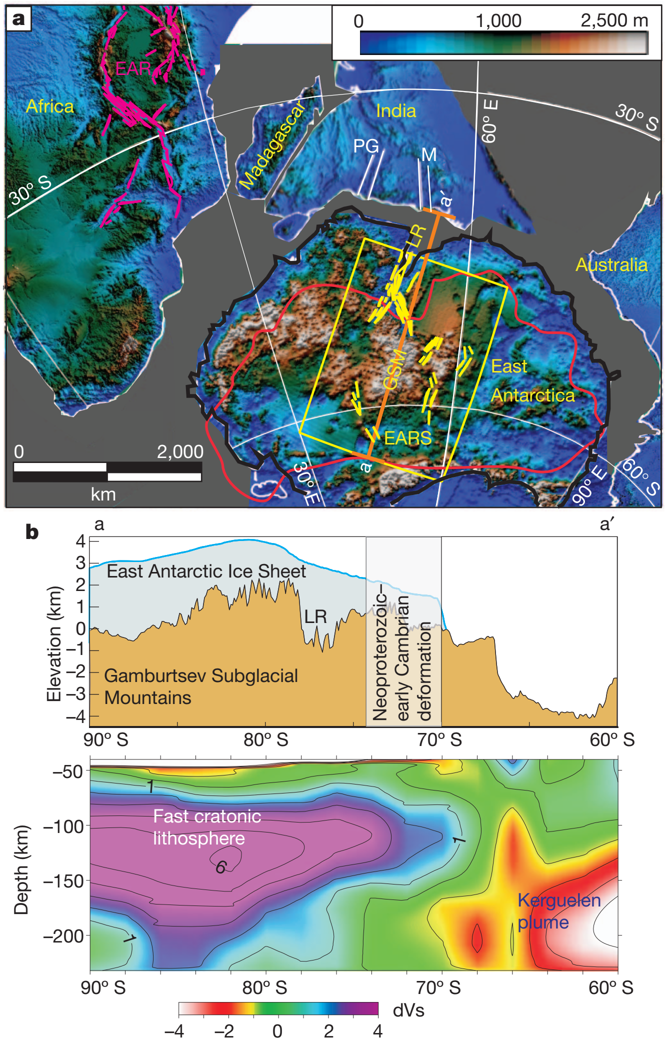


Figure 1 | East Antarctica in Gondwana and seismic tomography through the Gamburstev Subglacial Mountains. a, Gondwana at 140 Myr reconstructed using GPlates software (see Methods), including palinspastic restoration of Antarctic and Indian margins[30] and isostatically rebounded (that is, after removal of the ice sheet; thick black line) Antarctic topography. Yellow rectangle, study area. Orange line, location of the seismic tomography model shown in **b**. Red line, extent of fast cratonic lithosphere in East Antarctica at 150 km depth[10]. Yellow lines, East Antarctic rift system (EARS), including the Lambert rift (LR) and wrapping around the Gamburtsevs (GSM). Pink lines, East African rift system (EAR), which has a similar geometry to the EARS. White lines, Permian–Cretaceous Mahanadi (M) and Pranhita-Godavari (PG) rifts in India[23,30], conjugate to the EARS. **b**, Regional seismic tomography model along 75° E (present-day coordinates) through the Gamburtsevs and the Lambert rift. Contour lines show percentage variations in seismic shear-wave speed (dVs) with respect to the preliminary Earth reference model[10]. The thick, high-velocity cratonic lithosphere beneath the Gamburtsevs lies south of the region affected by Neoproterozoic–early Cambrian deformation associated with Gondwana assembly[8].

[1]British Antarctic Survey, High Cross, Madingley Road, Cambridge, CB3 0ET, UK. [2]US Geological Survey, Denver, Colorado 80225, USA. [3]Lamont Doherty Earth Observatory of Columbia University, Palisades, New York 10964, USA. [4]Bundensanstalt für Geowissenschaften und Rohstoffe, Hannover, Stilleweg 2, 30655, Germany.

120,000 km of new aerogeophysical data were collected (Supplementary Fig. 1), and are merged here (Fig. 2) with satellite[14,15] and previous airborne data[16–18], providing the most comprehensive geophysical perspective so far of crustal architecture and mountain-building processes in interior East Antarctica.

The new radar data reveal highly dissected Alpine topography reaching maximum elevations of 3,000 m and a median elevation of about 1,400 m in the north–south-trending Gamburtsevs (Fig. 2a). The aero- and satellite-magnetic data distinguish several basement provinces: the Archaean Ruker[16], Gamburtsev, Vostok[18], South Pole and Recovery, indicating that interior East Antarctica is probably composed of a mosaic of different Precambrian cratons and orogens (Fig. 2b and Supplementary Figs 2–4). The linear, northeast-trending magnetic anomalies that characterize much of the Gamburtsevs are similar to anomalies associated with Precambrian sources such as roughly 1-Gyr granites and orthogneisses to the north[16,17] (Fig. 2b). Regional northeast-trending Bouguer gravity lows (Fig. 2c) define thickened Precambrian crust under the Gamburtsev province. We calculated the effective elastic thickness of the lithosphere (T_e)[19], a proxy for its long-term strength, from three-dimensional (3D) inversion (Supplementary Figs 8–10). Our models indicate that the Gamburtsevs exhibit high T_e (more than 70 km), which is typical of Precambrian cratonic provinces[19] (Fig. 2d).

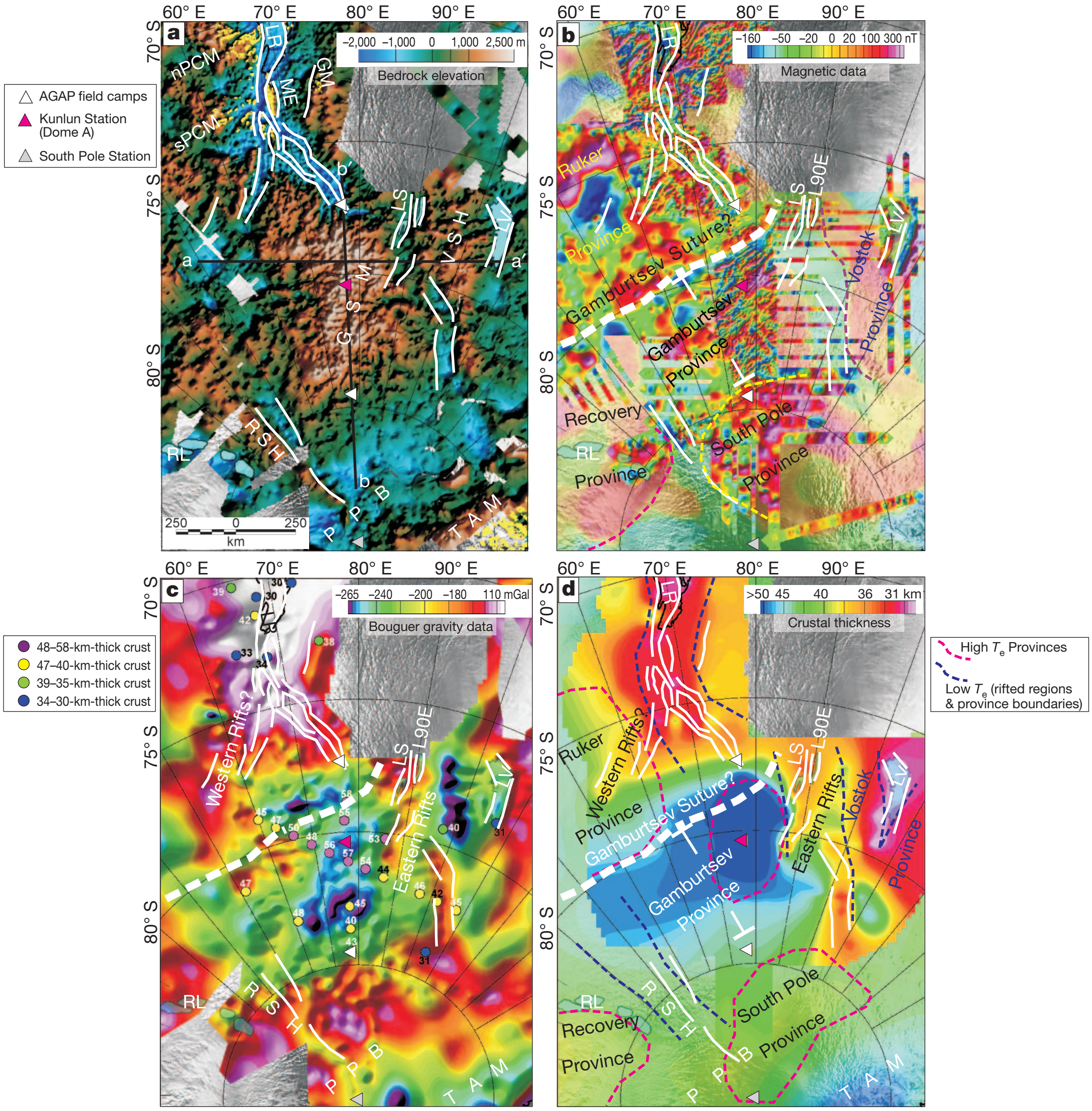


Figure 2 | Geophysical data and interpretation over central East Antarctica. **a**, Bedrock elevation data for the Gamburtsevs merged with BEDMAP (http://www.antarctica.ac.uk/bedmap/) and other existing radar data[17,18]. White lines, rift basins of the East Antarctic rift system. Black lines, locations of the models shown in Fig. 3. Yellow, outcrops. Light blue with black outlines, major subglacial lakes (LV, Lake Vostok; LS, Lake Sovetskaya; L90E, Lake 90 East; RL, Recovery Lakes). nPCM, northern Prince Charles Mountains; sPCM, southern Prince Charles Mountains; GM, Grove Mountains; ME, Mawson Escarpment; LR, Lambert rift; GSM, Gamburtsev Subglacial Mountains; VSH, Vostok Subglacial Highlands; PPB, Pensacola–Pole basin; RSH, Resolution Subglacial Highlands; TAM, Transantarctic Mountains. **b**, Aeromagnetic data for the Gamburtsevs merged with existing airborne (bright colours)[16,17] and satellite[14] (background) magnetic data. Dashed lines show Precambrian province boundaries in central East Antarctica inferred from magnetic anomalies. The inferred Gamburtsev suture separates the Archaean Ruker province[16] from the Proterozoic[21] Gamburtsev province. The eastern rifts developed along the boundary between the Gamburtsev and Vostok[18,21] provinces. The southwestern rifts probably exploited the boundary between the Recovery and South Pole provinces. **c**, Bouguer gravity data for the Gamburtsevs merged with existing airborne[17] and satellite[15] gravity data. Dots with numbers refer to crustal thickness (in km) derived from seismic receiver functions[4,20]. Anomalously thick crust beneath the Gamburtsev province is interpreted as linked to Proterozoic (roughly 1 Gyr) collisional process. **d**, Crustal-thickness map calculated from 2D gravity modelling and seismic receiver functions, with superimposed regions of high and low T_e derived from 3D flexural modelling. The Lambert, Eastern and Vostok rifts are characterized by thinned crust and weaker lithosphere. The Gamburtsev, Ruker, Recovery and South Pole provinces have stronger lithosphere.

A crust roughly 40 km thick[20] underlies Precambrian provinces in the Prince Charles Mountains (Fig. 2a). These provinces were deformed during Neoproterozoic–early Cambrian events (550–490 Myr) attributed to collisional or intraplate events associated with Gondwana assembly[8] (Fig. 1b). Seismological data indicate that beneath the Gamburtsevs, both the Precambrian crust and the lithosphere are much thicker (about 45–58 km thick[4] and 200 km thick[10] respectively). The gravity data require that the 12–18-km-thick root[4] under the northern and central Gamburtsevs is anomalously high density in its lower part, with a density contrast of only about 55 kg m^{-3} with the mantle (Fig. 3a, b and Supplementary Figs 5–7). These high densities are comparable to those modelled beneath old collisional orogens whose dense lower crustal roots lost buoyancy owing to temperature decreases and metamorphism[5,6]. Palaeoproterozoic (1.8–1.6 Gyr) and Mesoproterozoic–Neoproterozoic (1.2–0.8 Gyr) ages of the Gamburtsev and Vostok provinces, based on detrital zircons from the Vostok ice cores[21], suggest that collisional events associated with Rodinia or earlier supercontinental assembly formed the crustal root (Fig. 4a). A more recent age of 550–490 Myr for the Gamburtsevs' root is less likely, given that the roots of most coeval orogens linked with Gondwana amalgamation delaminated into the mantle[22]. The crustal root is preserved today in the same way as in the Palaeoproterozoic (roughly 1.8 Gyr) Trans-Hudson orogen and the Palaeozoic (roughly 300 Myr) Ural Mountains[6], but the ancestral Gamburtsev Mountains formed during the inferred Proterozoic collision would not have survived erosion for more than a few hundred million years (Fig. 4b).

More recent tectonic events must have rejuvenated the crustal root and triggered uplift of the modern Gamburtsevs. The Gamburtsev province was regionally elevated ground at the focus of radial drainage starting from about 300 Myr[9]. Permian extension (around 250 Myr) and Cretaceous strike-slip faulting (roughly 100 Myr) followed, forming the Lambert and conjugate rifts in India[23] (Fig. 1a). Crustal thinning beneath the northern Lambert rift is imaged by both seismic[20] and gravity data[17] (Fig. 2c, d and Supplementary Fig. 4). Our new data reveal that the Lambert rift branches at the northwestern edge of the Gamburtsevs (around the Archaean Ruker province) into narrow basins that surround the Proterozoic Gamburtsev province (Fig. 2b). The eastern branch includes a series of eastward-stepping rifts that separate the Gamburtsevs from the Vostok Subglacial Highlands (Fig. 2a). Collectively, the rifts form part of the newly identified East Antarctic rift system, with similar geometry and length to the modern East African rift system[7] (Fig. 1a). Today, these rift basins host the largest of all subglacial lakes in Antarctica (Fig. 2a), which resemble in size rift lakes in East Africa. Bouguer gravity highs (Fig. 2c) and localized magnetic lows (Fig. 2b) characterize most of the rift basins flanking the Gamburtsevs. Gravity models and magnetic depth estimates indicate that the Lambert rift and the eastern basins contain at least 1–3 km of sediment on top of thinned Precambrian crust (Fig. 3a, b). The eastern rift basins formed along the lithospheric boundary between the Gamburtsev and Vostok provinces, which includes an inferred Proterozoic foreland sedimentary basin[18] (Fig. 3a). The East Antarctic rift system features low T_e values ($25 < T_e < 40$ km) (Fig. 2d and Supplementary Fig. 10), probably reflecting pre-existing zones of weakness separating the Archaean–Proterozoic provinces, similar to those in East Africa[7] and other regions[19].

Extension within these rift basins, the buoyancy of the crustal root and fluvial and glacial erosion all contributed to Gamburtsev uplift. Mechanical unloading of the lithosphere during extension and consequent isostatic rebound produces high topography along rift flanks[24]. Rifting and exhumation resulted in 3–5 km of Permian–Jurassic denudation and 2–4.5 km of late Cretaceous–early Palaeocene denudation in the Prince Charles Mountains and Mawson Escarpment[25], flanking the Lambert rift (Fig. 2a). To evaluate the contribution of each mechanism to the Gamburtsev topography, we modelled the flexural

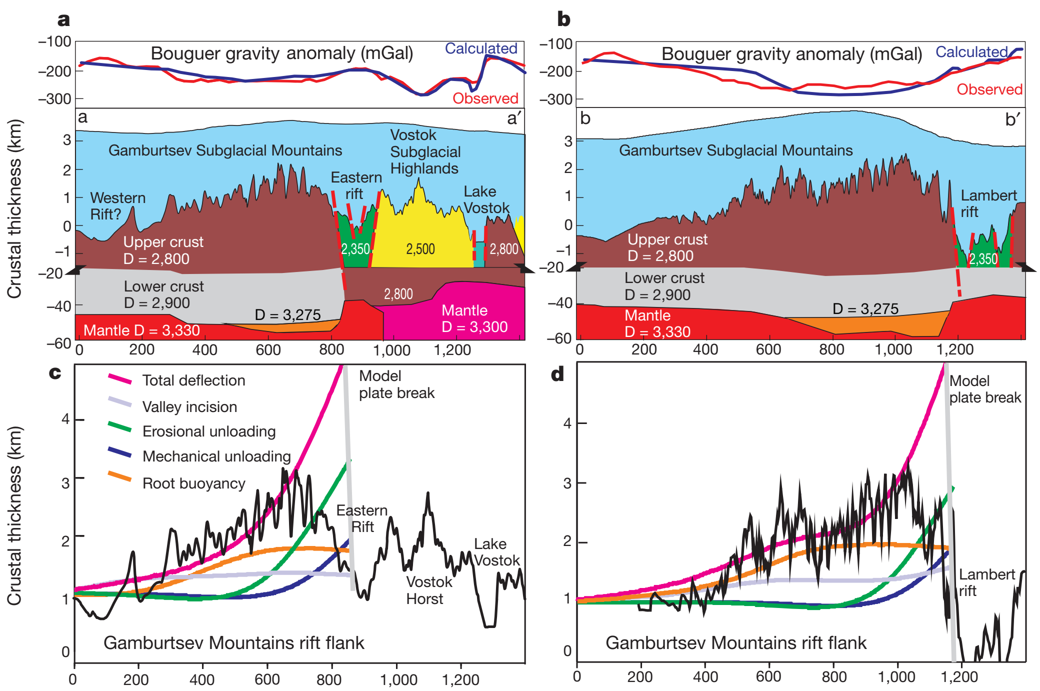


Figure 3 | Crustal architecture and uplift processes for the Gamburtsev Subglacial Mountains. **a**, **b**, Gravity models constrained by magnetic depth estimates and seismic crustal-thickness estimates[4] along an east–west (**a**) and north–south (**b**) profile. D, densities in kg m^{-3}. The root under the Gamburtsevs has two densities: 2,900 and 3,275 kg m^{-3} (orange). Eastern rift and Lambert rift sediments shown in green. The inferred Proterozoic foreland basin sediments[18,21] of the Vostok Subglacial Highlands are shown in yellow. Note significant differences in crust and mantle densities beneath the Gamburtsev and Vostok provinces. **c**, **d**, Results of flexural modelling along the east–west (**c**) and north–south (**b**) profile, assuming a T_e of 80 km and a broken-plate approximation. The crustal root contribution was modelled as a sum of the two lower crustal bodies. The response to incision and denudation was calculated using a rock density of 2,800 kg m^{-3} and a mantle density of 3,330 kg m^{-3}. The combination of root buoyancy, mechanical and erosional unloading along the rift flank and the isostatic response to valley incision drives Gamburtsev Mountains uplift (see also Supplementary Fig. 11).

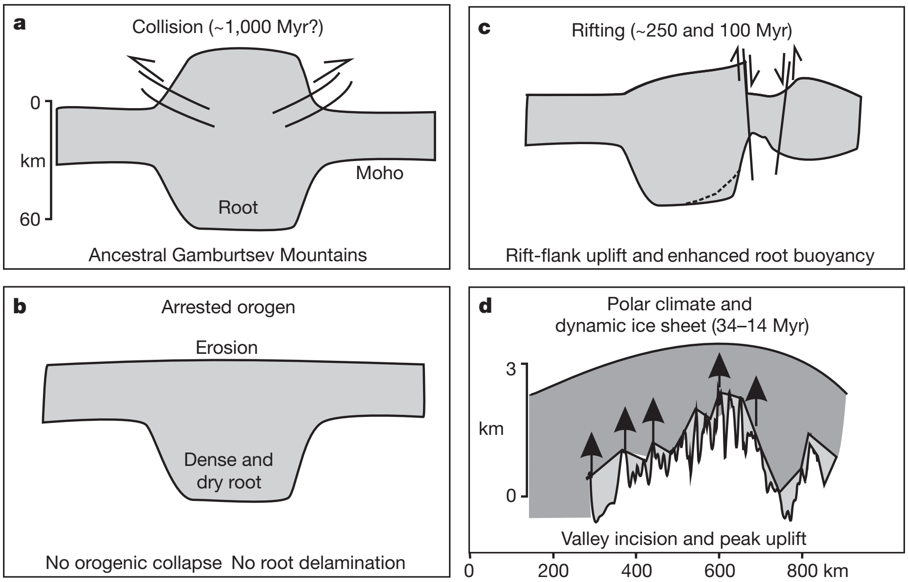


Figure 4 | Schematic of the elements contributing to Gamburtsev Mountains uplift. **a**, Proterozoic collision (about 1 Gyr) forms thick root and high topography of the ancestral Gamburtsev Mountains. **b**, Post-collisional orogenic collapse and major Moho re-equilibration did not occur, preserving a dry and dense root like that in the Trans-Hudson orogen and the Urals[5,6]. The elevated topography of the ancestral Gamburtsev Mountains was eroded. **c**, Permian (roughly 250 Myr) and Cretaceous (roughly 100 Myr) rifting drove flexural uplift and, through possible heating and/or depressurization[26], reduced the density of the root and released its latent buoyancy, to help produce the broad Gamburtsev rift-flank. **d**, Fluvial and glacial erosion in the valleys uplifted the peaks, creating the modern high-relief Alpine topography[2] of the Gamburtsevs. The East Antarctic Ice Sheet has preserved the rugged topography of the Gamburtsevs for at least 14 Myr (ref. 2).

response of the Gamburtsev lithosphere to mechanical unloading, erosional unloading, the crustal root and valley incision (Fig. 3c, d and Supplementary Fig. 11). The large gradient in T_e between the weaker rift basins and the stronger Gamburtsev province (Fig. 2d) suggests that a broken elastic plate is a good approximation of the lithospheric structure. With a broken-plate model, mechanical unloading generates around 1 km of uplift at the fault edge, falling to less than 200 m at around 200 km inboard. The model assumes that roughly 4 km of rock has been removed as a result of denudation[25], giving rise to about a further 2 km of uplift (Fig. 3c, d). Petrologic modelling of crustal roots suggests that increasing temperatures and decreasing pressure[26], such as would be expected during rifting and rift-flank uplift, may enhance root buoyancy (Fig. 4c). The buoyancy of the crustal root beneath the Gamburtsevs has a major impact on rift-flank uplift (Fig. 3c, d), because it increases elevation by about 1 km over a 500–700-km-wide region. In polar climates, valley incision can produce significant extra isostatic peak uplift[24], and the ice sheet can help to preserve Alpine topography[2] (Fig. 4d). The response to valley incision is around 500 m of uplift within the Gamburtsevs (Fig. 3c, d). Together, these tectonic, erosional and isostatic processes can explain the high elevation and relief of the modern Gamburtsevs (Figs 3c, d and 4c, d).

The location of the Gamburtsevs within the mosaic of Precambrian provinces in central East Antarctica is an important control on its mechanical properties and subsequent response to continental rifting. The Permian and Cretaceous East Antarctic rift system developed around the strong Archaean Ruker and Proterozoic[21] Gamburtsev provinces, in much the same way as Proterozoic orogens localized the modern East African rift system around the stronger Tanzania craton[7]. In East Africa, the absence of thick crustal roots beneath the Proterozoic orogens and the Archaean Tanzania craton[7] resulted in narrow rift-flank uplift. In interior East Antarctica, the broader rift flank is supported by a strong lower-crustal root, probably composed of dry granulite, as is inferred over the formerly adjacent (Fig. 1a) Indian shield[27]. The preservation of apparently intact and thick Precambrian lithosphere beneath the Gamburtsevs (Fig. 1b) differs from other continental interiors, where recent mountain uplift is associated with significant destabilization in the underlying lithosphere (for example, in the Colorado Plateau[28]). The broad and long-lived uplift that we propose in central East Antarctica provided a key nucleation site for the continental-scale Antarctic ice sheet around 34 Myr (ref. 2), and probably also for smaller late Cretaceous to early Cenozoic (about 70–50 Myr) ephemeral ice sheets[29].

METHODS SUMMARY

We combined AGAP aerogeophysical data with adjacent data sets and recent satellite magnetic and gravity data to investigate the crustal structure and origin of the enigmatic Gamburtsev Subglacial Mountains. Magnetic data provided the tool to reveal a mosaic of Precambrian basement provinces in interior East Antarctica. Ten two-dimensional (2D) forward gravity models were calculated to estimate crust–mantle-boundary (Moho) depth and crustal and upper-mantle density structure. We tied our gravity models to independent receiver-function estimates of crustal thickness[4,20] and isostatic models to reduce the inherent ambiguities; they incorporated depth-to-magnetic-source calculations from Werner deconvolution to help model the extent of sedimentary basins. We reconciled large misfits between gravity, flexural and seismic models by including an anomalously high-density lower crust beneath the Gamburtsevs. We derived an elastic-thickness grid (T_e) for interior East Antarctica from 3D inversion, using a spatial convolution method and taking into account both surface and intracrustal loads. We used a 2D finite difference method to model the flexural uplift of the Gamburtsevs with a broken-plate approximation, by incorporating the buoyancy of the crustal root, the effects of mechanical and erosional unloading along the flanks of the East Antarctic rift system, and the isostatic response to valley incision.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 May; accepted 16 September 2011.

1. Sorokhtin, O., Avsyuk, G. Y. & Koptev, V. I. Determination of the thickness of the ice cap in East Antarctica *Inform. Bull. Soviet Antarctic Expedition* **11,** 9–13 (1959).
2. Bo, S. *et al.* The Gamburtsev Mountains and the origin and early evolution of the Antarctic Ice Sheet. *Nature* **459,** 690–693 (2009).
3. Cox, S. E. *et al.* Extremely low long-term erosion rates around the Gamburtsev Mountains in interior East Antarctica. *Geophys. Res. Lett.* **37,** L22307 (2010).
4. Hansen, S. E. *et al.* Crustal structure of the Gamburtsev Mountains, East Antarctica, from S-wave receiver functions and Rayleigh wave phase velocities. *Earth Planet. Sci. Lett.* **300,** 395–401 (2010).
5. Fischer, K. M. Waning buoyancy in the crustal roots of old mountains. *Nature* **417,** 933–936 (2002).
6. Leech, M. L. Arrested orogenic development: eclogitization, delamination and tectonic collapse. *Earth Planet. Sci. Lett.* **185,** 149–159 (2001).

7. Petit, C. & Ebinger, C. Flexure and mechanical behavior of cratonic lithosphere: gravity models of the East African and Baikal rifts. *J. Geophys. Res.* **105,** 19151–19162 (2000).
8. Boger, S. D. *et al.* Pan-African intraplate deformation in the northern Prince Charles Mountains, East Antarctica. *Earth Planet. Sci. Lett.* **195,** 195–210 (2002).
9. Veevers, J. J., Saeed, A. & O'Brien, P. E. Provenance of the Gamburtsev Subglacial Mountains from U–Pb and Hf analysis of detrital zircons in Cretaceous to Quaternary sediments in Prydz Bay and beneath the Amery Ice Shelf. *Sedim. Geol.* **211,** 12–13 (2008).
10. Ritzwoller, M. H., Shapiro, N. M., Levshin, A. L. & Leahy, G. M. Crustal and upper mantle structure beneath Antarctica and surrounding oceans. *J. Geophys. Res.* **106,** 30,645–30,670 (2001).
11. Fitzsimons, I. C. W. in *Proterozoic East Gondwana: Supercontinent Assembly and Breakup* (eds Yoshida, M., Windley, B. F. & Dasgupta, S.) 93–103 (Geol. Soc. London, 2003).
12. Veevers, J. J. Case for the Gamburtsev Subglacial Mountains of East Antarctica orginating by mid-Carboniferous shortening of an intracratonic basement. *Geology* **22,** 593–596 (1994).
13. Sleep, N. H. Mantle plumes from top to bottom. *Earth Sci. Rev.* **77,** 231–271 (2006).
14. Maus, S. *et al.* Resolution of direction of oceanic magnetic lineations by the sixth-generation lithospheric magnetic field model from CHAMP satellite magnetic measurements. *Geochem. Geophys. Geosyst.* **9,** Q07021 (2008).
15. Pail, R. *et al.* Combined satellite gravity field model GOCO01S derived from GOCE and GRACE. *Geophys. Res. Lett.* **37,** L20314 (2010).
16. Golynsky, A. V., Alyavdin, S. V., Masolov, V. N., Tscherinov, A. S. & Volnukhin, V. S. The composite magnetic anomaly map of the East Antarctic. *Tectonophysics* **347,** 109–120 (2002).
17. McLean, M. A. *et al.* Basement interpretations from airborne magnetic and gravity data over the Lambert Rift region of East Antarctica. *J. Geophys. Res.* **144,** B06101 (2009).
18. Studinger, M. *et al.* Geophysical models for the tectonic framework of the Lake Vostok region, East Antarctica. *Earth Planet. Sci. Lett.* **216,** 663–677 (2003).
19. Audet, P. & Bürgmann, R. Dominant role of tectonic inheritance in supercontinent cycles. *Nature Geosci.* **4,** 184–187 (2011).
20. Reading, A. M. The seismic structure of Precambrian and early Palaeozoic terranes in the Lambert Glacier region, East Antarctica. *Earth Planet. Sci. Lett.* **244,** 44–57 (2006).
21. Leitchenkov, G. L., Belyatsky, B. V., Rodionov, N. V. & Sergeev, S. A. in *Antarctica: A Keystone in a Changing World — Online Proceedings of the 10th ISAES* (eds Cooper, A. K. & Raymond, C. R.) Short res. paper 14 (USGS Open-File Report 2007–1047, 2007).
22. Avigad, D. & Gvirtzman, Z. Late Neoproterozoic rise and fall of the northern Arabian–Nubian shield: The role of lithospheric mantle delamination and subsequent thermal subsidence. *Tectonophysics* **477,** 217–228 (2009).
23. Phillips, G. & Läufer, A. L. Brittle deformation relating to the Carboniferous–Cretaceous evolution of the Lambert Graben, East Antarctica: A precursor for Cenozoic relief development in an intraplate and glaciated region. *Tectonophysics* **471,** 216–224 (2009).
24. Stern, T. A., Baxter, A. K. & Barrett, P. J. Isostatic rebound due to glacial erosion within the Transantarctic Mountains. *Geology* **33,** 221–224 (2005).
25. Lisker, F., Gibson, H., Wilson, C. J. & Läufer, A. in *Antarctica: A Keystone in a Changing World – Online Proceedings of the 10th ISAES* (eds Cooper, A. K. & Raymond, C. R.) Short res. paper 105 (USGS Open-File Report 2007–1047, 2007).
26. Semprich, J., Simon, N. S. C. & Podladchikov, Y. Y. Density variations in the thickened crust as a function of pressure, temperature, and composition. *Int. J. Earth Sci.* **99,** 1487–1510 (2010).
27. Jackson, J. A., Austrheim, H., McKenzie, D. & Priestley, K. Metastability, mechanical strength, and the support of mountain belts. *Geology* **32,** 625–628 (2004).
28. Levander, A. *et al.* Continuing Colorado plateau uplift by delamination-style convective lithospheric downwelling. *Nature* **472,** 461–465 (2011).
29. Miller, K. G. *et al.* The Phanerozoic record of global sea-level change. *Science* **310,** 1293–1298 (2005).
30. Veevers, J. J. Palinspastic (pre-drift and –drift) fit of India and conjugate Antarctica and geological connections across the suture. *Gondwana Res.* **16,** 90–108 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the seven nations involved in the AGAP International Polar Year effort for their major logistical, financial and intellectual support. The US Antarctic Program of the National Science Foundation provided support for the logistics, the development of the instrumentation and data analysis. The Natural Environment Research Council/British Antarctic Survey provided support for deep-field operations, data collection and analysis. The Federal Institute for Geosciences and Resources provided financial support. The Australian Antarctic Division provided support at the AGAP North field camp; the Chinese Antarctic programme and the Alfred Wegner Institute also assisted. We thank all the AGAP project members involved, and in particular M. Studinger, N. Frearson and C. Robinson. C. Ebinger provided an early review and P. Molnar provided discussions. S. Golynsky provided geophysical data over adjacent regions and related discussions. We thank C. Braitenberg for assistance with Lithoflex and R. Buck for providing 2D flexural modelling code. J. J. Veevers provided a review.

Author Contributions F.F. processed magnetic data, compiled radar, magnetic and gravity images, performed gravity modelling and, with C.A.F., led data interpretation and paper development. T.A.J. processed the gravity data and ran the 2D flexural models. R.E.B. helped in writing sections of the paper. L.M.A. performed elastic thickness modelling and Gondwana reconstruction. D.D. contributed magnetic data processing. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to F.F. (ffe@bas.ac.uk).

METHODS

Bedrock topography compilation and plate reconstruction model. To analyse the Gamburtsev Subglacial Mountains, we first compiled a new bedrock topography grid for East Antarctica (Figs 1a and 2a) by merging the AGAP radar data with the BEDMAP compilation[31] (including Russian line data)[32] and more recent data over Vostok[18] and the southern Lambert rift[33] (Supplementary Fig. 1). The merged data were gridded using a kriging algorithm[34] with a grid mesh of 2,500 m. We calculated the isostatic correction associated with the removal of the overlying East Antarctic Ice Sheet, assuming a high effective elastic thickness ($T_e = 80$ km) for East Antarctica. The Antarctic rebounded-topography grid was then inputted into GPlates reconstruction software (http://www.gplates.org/), together with down-sampled (1-km resolution) digital elevation models for Africa, India and Australia from the Shuttle Radar Topography Mission (http://srtm.usgs.gov/). To derive the Gondwana reconstruction at 140 Myr (Fig. 1a), we used the Global EarthByte Rotation File[35] released with GPlates version 0.9.10, based on moving Indian/Atlantic hotspots up to 100 Myr[36], and fixed Indian/Atlantic hotpots from 100–140 Myr[37]. We also included a recent palinspastic restoration of Antarctic and Indian margins that reveals the aligned Lambert and Mahanadi rifts in East Antarctica and India[30], which we interpret as part of the newly identified East Antarctic rift system (Fig. 1a).

Potential field data compilation. We used potential field data to image basement provinces and crustal structure. AGAP aeromagnetic data were microlevelled[38] and draped[39] onto the new subglacial topography at a distance 2.5 km above the bedrock, and were gridded with a minimum-curvature algorithm (1,250-m grid mesh). Adjacent Russian data[16]—included in the Antarctic Digital Magnetic Anomaly compilation[40]—and more recent aeromagnetic data from the Prince Charles Mountains Expedition of Germany–Australia[17] and Vostok[18] (Supplementary Fig. 1) were also microlevelled and draped at the same elevation. We then stitched the individual grids together in frequency domain[41]. To help unveil province boundaries, we applied a terracing filter[42] to the MF6 satellite-derived global magnetic model[14] (continued at the same elevation as the airborne data), and backdropped the resulting grid beneath the aeromagnetic data (Fig. 2b). The satellite data are effective in delineating some large-scale (greater than 500-km) basement provinces in interior East Antarctica (as in other continents[43]), but cannot resolve the shallower crustal architecture imaged from aeromagnetic data (Supplementary Figs 2–3). We observed a broad satellite low in the Gamburtsev Province, but did not resolve the higher-frequency northeast–southwest trends mapped from aeromagnetic data (Supplementary Fig. 2). We applied Werner deconvolution techniques[44] to the aeromagnetic data to estimate depth to magnetic sources as a proxy for sedimentary infill along our gravity models (Fig. 3 and Supplementary Figs 5–6).

We used Parker inversion[45] to derive the complete Bouguer gravity anomalies for the Gamburtsevs by applying a 3D correction for ice thickness and bedrock topography to AGAP Free-Air airborne gravity data. We assumed bulk densities of 910 kg m^{-3} and 2,670 kg m^{-3} for ice and bedrock respectively, and continued all data to a common elevation of 4,600 m (corresponding to the highest flight level). We gridded the airborne data with a minimum-curvature algorithm using a grid mesh of 1,250 m, and we subsequently low-pass filtered the data with a wavelength of 50 km to emphasize deeper sources. We then merged the airborne gravity data with existing airborne gravity data sets[17,18] and with long-wavelength (more than 200 km) satellite-derived Bouguer gravity anomalies for East Antarctica (Fig. 2c and Supplementary Fig. 4a, c) (calculated from the global gravity model *GOCO01S*[15]). The merged grid forms the basis for our gravity modelling across the Gamburtsevs and the East Antarctic rift system.

Gravity and flexural modelling. We computed ten 2D forward gravity models to image crustal architecture and depth to Moho beneath the Gamburtsevs and the East Antarctic rift system (Supplementary Fig. 4b, c). To reduce the ambiguities associated with gravity modelling, we tied our models to independent receiver-function analysis[4,20] and incorporated flexural isostatic compensation of the subglacial topography and intracrustal loads[46]. Initial forward gravity models, calculated assuming a standard density contrast of 500–300 kg m^{-3} between the crust and the mantle, yielded a consistently shallower Moho beneath the Gamburtsevs than did receiver-function estimates[4] (about 40–45 km instead of 50–58 km). 3D inversion of satellite gravity data confirmed the misfit, and flexural models incorporating surface loads alone also failed to reproduce the seismically observed Moho and root geometry. We found that depth-to-Moho estimates from gravity and receiver functions can be reconciled by assuming that an anomalously high-density lower crustal root underlies the northern and central Gamburtsevs (Supplementary Figs 5–7). We obtained the best fit for apparent densities of 3,275 kg m^{-3}, although relatively lower root densities of 3,200 kg m^{-3} may be permissible, even without assuming anomalously high upper crustal densities (Supplementary Table 1). Such high apparent densities in the lower crust exceed those derived from gravity models of the dense Uralian root, where values of 3,100 kg m^{-3} have been suggested to reveal mafic granulites[47]. Comparable values have however been modelled for the roots of some Proterozoic orogens (older than 1 Gyr), including the northwest Grenville orogen, Trans-Hudson orogen and Svecofennian orogen[5], and would fall, for example, within the density range of garnet-bearing mafic granulites (3,000–3,300 kg m^{-3}; ref. 48).

The magnitude of T_e and its spatial variations exert a significant effect on the degree and style of deformation, and hence response to long-term surface and intracrustal loads[19,46,49]. Estimating T_e is therefore an important component in our uplift model for the Gamburtsevs. The traditional methodology of estimating T_e involves using admittance and/or coherency techniques with free-air or Bouguer-gravity-anomaly data respectively, and has led to contrasting results, in particular over cratonic regions[49,50]. Inversion methods based on the spatial convolution of surface and crustal loads offer an alternative approach that can provide higher spatial resolution[51], and are used in our study. The 3D inversion was run using the Lithoflex software (http://www.lithoflex.org/) that has previously been used to estimate T_e variations for several mountain ranges, including the Alps and the Andes[52,53]. Using the rebounded bedrock topography, the equivalent topography of the dense lower crustal root and the modelled Moho grid as inputs, we obtained a grid of estimated T_e variations for the Gamburtsevs and adjacent East Antarctic rift system (Fig. 2d and Supplementary Figs 8–10). Our models indicate that a strong cratonic crust and upper mantle (as recognized, for example, in the Canadian Shield[54], Tanzania Craton[7,19] and European Craton[50]) underlie the Gamburtsevs, in contrast with the weaker rifted crust surrounding the range.

Flexural models have been used widely to address rift-flank uplift processes[46], including those along the flanks of the East African rift system[7] and the Transantarctic Mountains[55]. We used a 2D finite-difference method[56] to model the flexural uplift of the Gamburtsevs rift flank (Fig. 3c,d). We imposed distributed loads on a broken elastic plate with a T_e of 80 km (consistent with high T_e values derived from 3D inversion) falling to zero at the plate breaks, corresponding to the margin of the Lambert and Eastern rifts (Supplementary Fig. 11). The modelled total deflection (Fig. 3c, d) is the sum of the root buoyancy, mechanical and erosional unloading, and local valley-incision effects. Our flexural model is a close match to the observed high elevation and relief of the Gamburtsev Subglacial Mountains in interior East Antarctica.

31. Lythe, M., Vaughan, D. G. & the BEDMAP Consortium. BEDMAP: a new ice thickness and subglacial topographic model of Antarctica. *J. Geophys. Res.* **106,** 11335–11351 (2001).
32. Popov, S. V. *et al. Antarctica: A Keystone in a Changing World — Online Proceedings of the 10th ISAES* (eds. Cooper, A. K. & Raymond, C. R.) Extended abstr. 26 (USGS Open-File Report 2007–1047, 2007).
33. Damm, V. A subglacial topographic model of the southern drainage area of the Lambert Glacier/Amery Ice Shelf System — results of an airborne ice thickness survey south of the Prince Charles Mountains. *Terra Antartica* **14,** 85–94 (2007).
34. Journel, A. G. & Huijbregts, C. J. *Mining Geostatistics* (Academic, 1978).
35. Müller, R. D., Sdrolias, M., Gaina, C. & Roest, W. R. Age, spreading rates, and spreading asymmetry of the world's ocean crust. *Geochem. Geophys. Geosyst.* **9,** Q04006 (2008).
36. O'Neill, C., Müller, R. D. & Steinberger, B. On the uncertainties in hot spot reconstructions and the significance of moving hot spot reference frames. *Geochem. Geophys. Geosyst.* **6,** Q04003 (2005).
37. Müller, R. D., Royer, J. L. & Lawver, L. A. Revised plate motions relative to the hotspots from combined Atlantic and Indian Ocean hotspot tracks. *Geology* **21,** 275–278 (1993).
38. Ferraccioli, F., Gambetta, M. & Bozzo, E. Microlevelling procedures applied to regional aeromagnetic data: an example from the Transantarctic Mountains (Antarctica). *Geophys. Prospect.* **46,** 177–196 (1998).
39. Pilkington, M. & Thurston, B. J. Draping corrections for aeromagnetic data: line versus grid-based approaches. *Explor. Geophys.* **32,** 95–101 (2001).
40. Golynsky, A. *et al.* in *BAS (Misc.)* (eds Morris, P. & von Frese, R.) Vol. 10 (British Antarctic Survey, 2001).
41. Johnson, A., Cheeseman, S. & Ferris, J. Improved compilation of Antarctic Peninsula magnetic data by new interactive grid suturing and blending methods. *Ann. Geofis.* **42,** 249–259 (1999).
42. Cooper, G. R. J. & Cowan, D. R. Terracing potential field data. *Geophysical Prospecting* **57,** 1067–1071 (2009).
43. Hemant, K. & Maus, S. Geological modeling of the new CHAMP magnetic anomaly maps using a geographical information system technique. *J. Geophys. Res.* **110,** B12103 (2005).
44. Ku, C. C. & Sharp, J. A. Werner deconvolution for automated magnetic interpretation and its refinement using Marquardt inverse modelling. *Geophysics* **48,** 754–774 (1983).
45. Parker, R. L. Rapid calculation of potential anomalies. *Geophys. J. R. Astron. Soc.* **31,** 447–455 (1973).
46. Watts, A. B. *Isostasy and Flexure of the Lithosphere.* (Cambridge University Press, 2001).
47. Scarrow, J. H., Ayala, C. & Kimbell, G. S. Insights into orogenesis: getting to the root of a continent-ocean-continent collision, Southern Urals, Russia. *J. Geol. Soc. Lond.* **159,** 659–671 (2002).

48. Rudnick, R. L. & Fountain, D. M. Nature and composition of the continental crust: a lower crustal perspective. *Rev. Geophys.* **33,** 267–309 (1995).
49. McKenzie, D. Estimating Te in the presence of internal loads. *J. Geophys. Res.* **108,** (2003).
50. Pérez-Gussinyé, M. & Watts, A. B. The long-term strength of Europe and its implications for plate-forming processes. *Nature* **436,** 381–384 (2005).
51. Wienecke, S., Braitenberg, C. & Goetze, H. J. A new analytical solution estimating the flexural rigidity in the Central Andes. *Geophys. J. Int.* **169,** 789–794 (2007).
52. Braitenberg, C., Ebbing, J. & Götze, H. J. Inverse modelling of elastic thickness by convolution method—the eastern Alps as a case example. *Earth Planet. Sci. Lett.* **202,** 387–404 (2002).
53. Gimenez, M. E., Braitenberg, C., Martinez, M. P. & Introcaso, A. A comparative analysis of seismological and gravimetric crustal thicknesses below the Andean Region with flat subduction of the Nazca Plate. *Int. J. Geophys.* **2009,** 607458 (2009).
54. Burov, E., Jaupart, C. & Mareshal, J. C. Large-scale crustal heterogeneities and lithospheric strength in cratons. *Earth Planet. Sci. Lett.* **164,** 205–219 (1998).
55. ten Brink, U. & Stern, T. Rift flank uplifts and hinterland basins: comparison of the Transantarctic Mountains with the Great Escarpment of Southern Africa. *J. Geophys. Res.* **97,** 569–585 (1992).
56. Karner, G. D. & Watts, A. B. Gravity anomalies and flexure of the lithosphere at mountain ranges. *J. Geophys. Res.* **88,** 10449–10477 (1983).

doi:10.1038/nature10516

Multiple routes to mammalian diversity

Chris Venditti[1], Andrew Meade[2] & Mark Pagel[2,3]

The radiation of the mammals provides a 165-million-year test case for evolutionary theories of how species occupy and then fill ecological niches. It is widely assumed that species often diverge rapidly early in their evolution, and that this is followed by a longer, drawn-out period of slower evolutionary fine-tuning as natural selection fits organisms into an increasingly occupied niche space[1,2]. But recent studies have hinted that the process may not be so simple[3–5]. Here we apply statistical methods that automatically detect temporal shifts in the rate of evolution through time to a comprehensive mammalian phylogeny[6] and data set[7] of body sizes of 3,185 extant species. Unexpectedly, the majority of mammal species, including two of the most speciose orders (Rodentia and Chiroptera), have no history of substantial and sustained increases in the rates of evolution. Instead, a subset of the mammals has experienced an explosive increase (between 10- and 52-fold) in the rate of evolution along the single branch leading to the common ancestor of their monophyletic group (for example Chiroptera), followed by a quick return to lower or background levels. The remaining species are a taxonomically diverse assemblage showing a significant, sustained increase or decrease in their rates of evolution. These results necessarily decouple morphological diversification from speciation and suggest that the processes that give rise to the morphological diversity of a class of animals are far more free to vary than previously considered. Niches do not seem to fill up, and diversity seems to arise whenever, wherever and at whatever rate it is advantageous.

Our approach uses a generalized least-squares model[8,9] of trait evolution in a Bayesian reversible-jump[10] framework that allows rates of evolution to vary in individual branches or entire monophyletic subgroups of a phylogeny (Supplementary Information). This allows us to trace the evolutionary history of shifts in the rate and timing of evolution without specifying in advance where these events are located, and to derive posterior probability density estimates of their magnitudes and probability of occurrence (Supplementary Information). The null model states that evolution has proceeded at a constant rate throughout the class Mammalia. Applied to log-transformed body size data ($n = 3{,}185$ species) arrayed on the mammalian tree[6], this model returns a Bayesian posterior density of log-likelihoods with a mean of -939.34 ± 0.99 (Fig. 1a), and a mean instantaneous rate of body size evolution of 1.02 g per million years. If rates are allowed to vary throughout the tree, the posterior density improves to a mean log-likelihood of -364.13 ± 23.01 (log(Bayes factor) = 993.51; values >10 considered 'very strong' support[11]; Fig. 1a). We detect evidence for a shift or change in the rate of evolution in approximately one-third, or 1,494 branches, of the tree, where to be included in this count branches had to either experience a change in rate in that branch or inherit that change from its immediate ancestral lineage, in at least 95% of the trees in the posterior sample. These shifts range from a 3-fold decrease to a 52-fold increase in the rate of evolution along a branch (Fig. 1b).

It has long been believed that the radiation of extant mammals underwent a burst of body-size evolution that occurred early in its history and coincided with the appearance of the mammalian orders, and that this was followed by a gradual slowdown towards the present[4,12–14]. Explanations for this pattern suppose that mammals moved into a largely unoccupied niche and geographical space as they came to be the dominant vertebrate group on Earth. Then, as time went on, niche space and unexplored geographical regions became scarce, reducing opportunities for diversification[4]. In striking contrast to this picture, we do not find any evidence for either a generalized burst of evolution early in mammalian evolution or for the rates of evolution to decrease as time moves towards the present (Fig. 2a). Instead, rates of evolution were low and stable for about the first 60 million years, only starting to increase around 90 million years ago and then showing only about a twofold increase over the previous 'baseline' rate. This increase occurred before the origin of the present-day mammalian orders and is

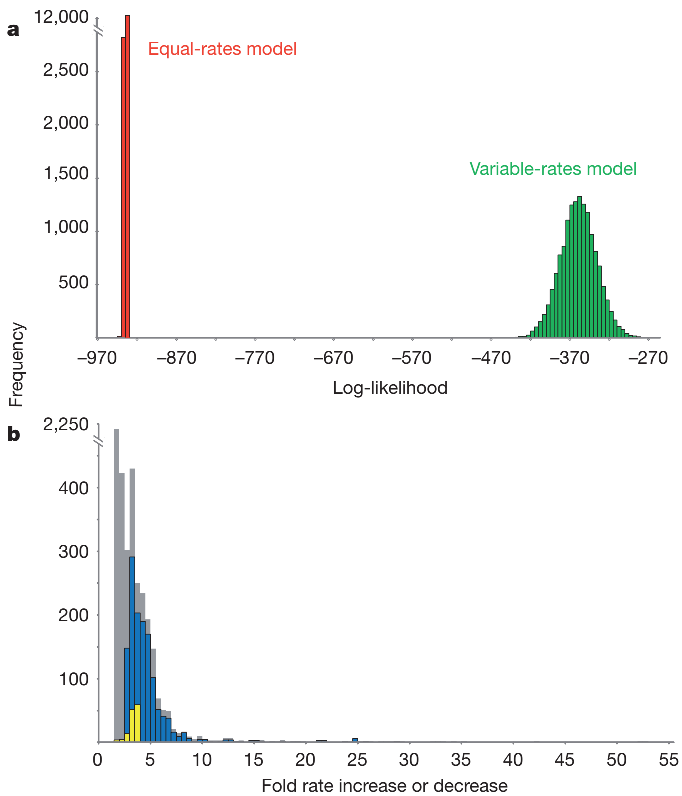


Figure 1 | Log-likelihood of trait models when rates are allowed to vary. **a**, Posterior distribution of log-likelihoods from a model with equal rates of evolution (red), compared with the posterior distribution of log-likelihoods from the model in which evolutionary rates are allowed to vary (green): log(Bayes factor) = 993.51 (calculated from the log-harmonic means of the likelihoods); values >10 considered 'very strong' support. **b**, The coloured bars show distributions of rates for the one-third of the branches (1,494) for which the posterior probability of having a rate shift was greater than 0.95. Blue bars signify *x*-fold rate increases and yellow bars indicate *x*-fold rate decreases. Grey bars show the distribution of the mean fold rates for all the branches in the mammal phylogeny, independent of the level of posterior support.

[1]Department of Biological Sciences, University of Hull, Hull HU6 7RX, UK. [2]School of Biological Sciences, University of Reading, Reading RG6 6BX, UK. [3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.

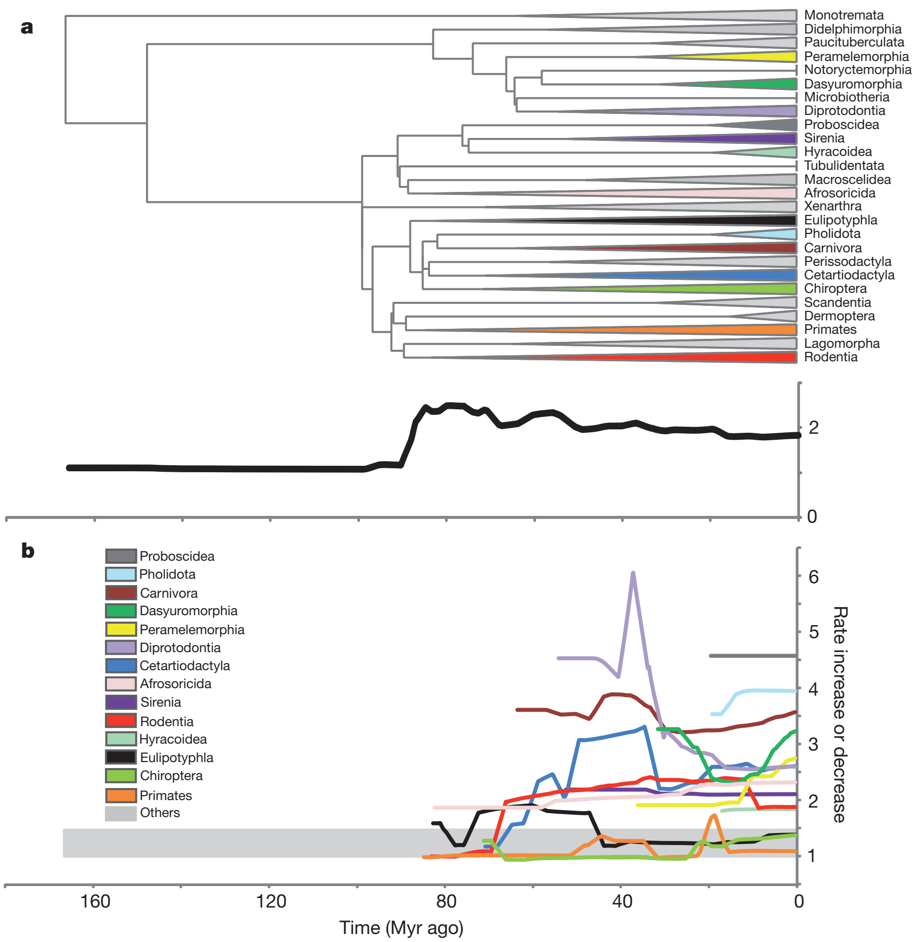


Figure 2 | Rates of mammalian morphological evolution through time. **a**, Mean time-dependent rates of evolution for the mammalian radiation taken as a whole. **b**, Mean time-dependent rates of evolution within each mammalian order. The colour key matches that in the dated phylogeny in **a** (dates taken from the corrigendum to ref. 6), which has been collapsed to the level of order; the start of each triangle indicates the first split in that order. Orders shaded grey in the phylogeny have rates throughout their evolutionary history that fall within the grey bar in **b**. The mean rates presented in **a** and **b** are calculated taking in to account the shared ancestry as implied by the phylogeny (mean rate decreases are less than one).

in accordance with estimates generated using a different technique[15]. Thus, rates of evolution increased at a point corresponding to the early splitting of the Laurasiatheria and the Afrotheria, and were then largely maintained until the present (Fig. 2a).

The time-dependent rates of change for the individual orders reveal a variety of patterns beginning around 80 million years ago (Fig. 2b). These tell three broad stories. The first is that there is no general tendency for orders to show bursts of evolution early in their radiations, followed by a gradual descent to baseline. In fact, in several orders—for example Primates, Eulipotyphla and Chiroptera—rates of change remain below the average mammalian rate (Fig. 2b). The second is that what might seem to be increases followed by decreases in a few orders, notably in Carnivora, Cetartiodactyla, Diprodontia and Rodentia, do not reflect order-wide processes. Rather, they are brought about by large changes in one or a small number of branches within those orders (Fig. 3 and Supplementary Figs 1–11). The third is that rates of body size evolution—and by implication many of the morphological and life history traits that allometrically scale with size[12]—are decoupled from speciation. The slowly evolving Chiroptera account for ~1,000 species, and the rate of evolution in Rodentia—the most speciose mammalian group, accounting for roughly 60% of all extant species—has rarely exceeded the mammalian trend (calculated rates account for the shared ancestry the tree implies and so are not biased by speciose clades). Instead, the highest average rates of change occur in one of the least speciose groups. The Proboscidea, some of whose members—the large elephants—are, along with the Sirenia, the closest living relatives to the small hyrax species, evolve on average 4.6-fold faster than the mammalian norm.

Looking more closely within orders, we see no evidence at any phylogenetic level for an 'early-burst' pattern of evolution. Instead, natural selection seems capable of altering body size spontaneously and over short periods of time. Thus, we observe short-term explosive increases of between 10- and 52-fold in the rate of evolution, distributed widely among clades (Fig. 3). It is these that shape the diversity of mammals, in striking ways. For example, the branches leading to Atelidae (woolly, spider and howler monkeys) record a 10-fold increase in the rate of evolution; the 'big cats' (the genera *Panthera*, *Acinonyx*, *Uncia* and *Puma*) jump to a 35-fold increase; Hominidae increase 12-fold; and Mysticeti increase 25-fold. The highest single-lineage burst, in excess of 52-fold, occurs in the branch leading to the genera *Dasyurus* (quolls) and *Sarcophilus* (Tasmanian Devils). We also observe stark differences in rate between closely related sister groups in many places across the mammal tree, including in the musk ox (*Ovibos moschatus*), which is much larger than its closest relatives, the gorals (*Naemorhedus*); and in the pygmy marmoset (*Callithrix pygmaea*), which is considerably smaller than its close relatives. The megabat genus *Taphozous* shows a drastic generalized decrease in the rate of evolution throughout the whole clade, whereas the genus *Dasyurus* and close relatives within the order Dasyuromorphia show a clear generalized increase (Supplementary Information). In a few cases, we observe a generalized increase in the rate of evolution throughout an entire clade (for example in Carnivora and Proboscidea). These

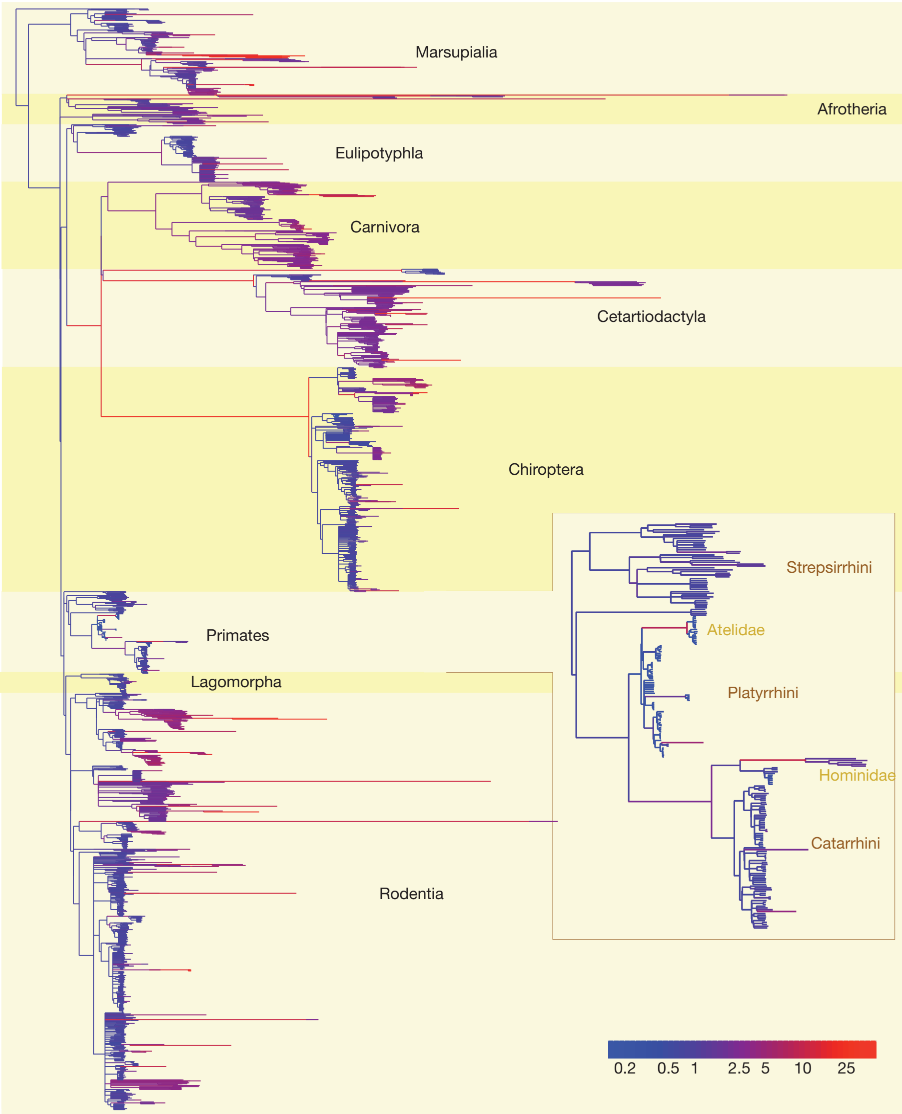


Figure 3 | The mammalian phylogenetic tree scaled to reflect morphological evolution. The branches of the phylogeny are transformed by the mean of the posterior distribution of the scalars acting on each branch—the branches of the tree are stretched and compressed to reflect the rate of morphological evolution. Also, the branches are coloured according to how much they have been scaled (scale factor shown in colour bar).

might reflect trends towards increased size such as described in some palaeontological and neontological studies[5,15,16].

An important subset of the short-term bursts occurs in the single evolutionary lineage that leads to the common ancestor of a large monophyletic group. These 'single-lineage ancestral bursts' seem to correspond to drastic changes in an animal from some ancestral state to a size that seems to form the basis for a new radiation, and might be akin to the concept of quantum evolution[14]. For example, the rate of evolution in the branch leading to Chiroptera is 24-fold higher than what we expect from a constant-rate model. This suggests that the common ancestor of the bats descended rapidly from a considerably larger, perhaps carnivore-, artiodactyl- or cetartiodactyl-like, ancestor, but that once it reached the 'bat' size, rates of change proceeded at background or even lower levels, even while the number of bat species increased dramatically. Other single-lineage ancestral bursts seem to characterize the changes leading to several orders in Laurasiatheria and Afrotheria (Table 1 and Fig. 3). These rate changes and those we observe more generally throughout the tree are large but are in line with those for some genes, or for some morphological traits whose changes have been measured in real time[17–19].

Conventional models that estimate a single homogeneous process over an entire clade—such as an early burst—can be easily misled by rate shifts that are confined to one or just a few branches. For example, recent studies have shown strong support for the early-burst model in primates[4] and cetaceans[20]. However, our results show that these bursts are attributable, in the case of primates, to two explosive increases along single branches (one leading to Atelidae and one to Hominidae; see inset in Fig. 3), and, in the cetaceans (the branch leading to Mysticeti; Supplementary Information), to one, rather than to a general trend. In other cases, a subgroup within a larger taxonomic range has increased its rate of change (for example the genera *Rhinolophus* in Chiroptera, *Peroryctes* in Peramelemorphia, and *Dasyurus* and *Sarcophilus* in Dasyuromorphia; Fig. 3 and Supplementary Information). This pattern

Table 1 | Fold-increase rate of evolution for mammalian orders whose ancestral lineage is increased more than tenfold

Order	Single-lineage rate*	Mean rate† in monophyletic group (s.d.)
Cetartiodactyla	18.7	3.6 (3.2)
Chiroptera	24.1	1.9 (1.8)
Perissodactyla	17.1	1.1 (0.1)
Proboscidea	14.4	4.6 (0.2)
Sirenia	15.2	2.2 (0.2)

* Rate increase in the single lineage leading to the corresponding order.
† Mean and s.d. of the of the rate increases along the branches within the corresponding order.

too has been wrongly assigned to a homogenous model, in this case to fitting an Ornstein–Uhlenbeck process across the entire order or clade (see, for example, ref. 4). It is possible that the early-burst pattern exists in our data but that our methods lack the power to detect it, but we think this unlikely for two reasons. One is that our methods can reliably detect twofold, or smaller, changes in the rates of evolution (Figs 1b and 2b; see also simulations in Supplementary Information). The second is that relaxing our statistical criteria to make it easier for changes to be identified does not change our conclusions (Supplementary Information). If early-burst patterns do exist in the mammals, they are so small as to be of little evolutionary significance in comparison to the striking shifts in the patterns of diversification we observe distributed throughout the tree.

We infer historical evolutionary patterns in the rate of body size evolution using a phylogeny constructed from extant species[8,9]. Incorporating fossils[3,21–24] in our analysis could reveal details of the patterns of evolution leading to extinct species, and these could even differ from those for extant species. For example, the rate of evolution in extinct groups such as the triconodonts and multituberculates, which fall between the extant monotremes and the rest of the extant mammals, could have a considerably higher (or lower) rate of change than extant groups. However, we would not expect the historical trends we describe here for extant species to change qualitatively with the addition of extinct groups, even if those extinct groups did show different patterns of change; and we would not expect the inclusion of fossils to produce generalized early-burst patterns of change. The reason is that the methodology we use is specifically designed to account for the sort of inhomogeneity that would occur if an extinct group had a distinct rate of change. Thus, if such a fossil group were included, our method would detect the difference in rate and scale in that section of the tree and leave the rates along the remainder of the phylogeny unchanged. The mammalian supertree[6] we use in this study, although for the most part well resolved, does contain some polytomies. Polytomies can artificially decrease the rate of evolution[4], but we do not believe that this greatly affects our results: the lack of resolution in the tree is mostly found at the tips, yet we find many instances of very high rate shifts in terminal branches. In fact, we find the most instances of high rates at the tips in the rodent clade, which contains the most polytomies.

Our results reveal that natural selection has been a precise and flexible shaper of mammalian size diversity, able to produce rapid changes in size in specific parts of the tree, and using a variety of 'substrates', be they elephants, carnivores, whales and even some rodents. Contrary to the long-held view that the diversity in mammal body sizes we see today is the product of widespread and homogeneous macroevolutionary processes in the rate of evolution, we find that natural selection has found multiple different routes to producing the current diversity of sizes. Our results also challenge the view that ecological niches fill up, as we find no suggestion of a generalized slowing in the rate of evolution as clades expand. This might suggest that ecological niches are a constantly 'moving target' and that they move just as much for speciose clades as for more depauperate ones. Combined with evidence that speciation rates might be constant in many groups of species, and that speciation might itself be the outcome of unusual single events[25], our results indicate that to understand so-called adaptive radiations it is necessary to study the multiple events in a species' life that provide it with the opportunity to adapt, rather than studying wide and general processes.

Received 4 May; accepted 30 August 2011.
Published online 19 October 2011.

1. Simpson, G. G. *Tempo and Mode in Evolution* (Columbia Univ. Press, 1944).
2. Foote, M. Morphological disparity in Ordovician-Devonian crinoids and the early saturation of morphological space. *Paleobiology* **20,** 320–344 (1994).
3. Harmon, L. J. *et al.* Early bursts of body size and shape evolution are rare in comparative data. *Evolution* **64,** 2385–2396 (2010).
4. Cooper, N. & Purvis, A. Body size evolution in mammals: complexity in tempo and mode. *Am. Nat.* **175,** 727–738 (2010).
5. Clauset, A. & Erwin, D. H. Evolution and distribution of species body size. *Science* **321,** 399–401 (2008).
6. Bininda-Emonds, O. R. P. *et al.* The delayed rise of present-day mammals. *Nature* **446,** 507–512 (2007); corrigendum **456,** 274 (2008).
7. Jones, K. E. *et al.* PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90,** 2648 (2009).
8. Pagel, M. Inferring evolutionary processes from phylogenies. *Zool. Scr.* **26,** 331–348 (1997).
9. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401,** 877–884 (1999).
10. Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82,** 711–732 (1995).
11. Raftery, A. E. in *Markov Chain Monte Carlo in Practice* (eds Gilks, W. R., Richardson, S. & Spiegelhalter, D. J.) 163–187 (Chapman & Hall, 1996).
12. Read, A. F. & Harvey, P. H. Life history differences among the eutherian radiations. *J. Zool.* **219,** 329–353 (1989).
13. Foote, M. Evolutionary patterns in the fossil record. *Evolution* **50,** 1–11 (1996).
14. Simpson, G. G. *The Major Features of Evolution* (Columbia Univ. Press, 1953).
15. Clauset, A. & Redner, S. Evolutionary model of species body mass diversification. *Phys. Rev. Lett.* **102,** 038103 (2009).
16. Alroy, J. Cope's rule and the dynamics of body mass evolution in North American Fossil mammals. *Science* **280,** 731–734 (1998).
17. Smith, F. A., Betancourt, J. L. & Brown, J. H. Evolution of body size in the woodrat over the past 25,000 years of climate change. *Science* **270,** 2012–2014 (1995).
18. Smith, F. A., Browning, H. & Shepherd, U. L. The influence of climate change on the body mass of woodrats Neotoma in an arid region of New Mexico, USA. *Ecography* **21,** 140–148 (1998).
19. Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443,** 167–172 (2006).
20. Slater, G. J., Price, S. A., Santini, F. & Alfaro, M. E. Diversity versus disparity and the radiation of modern cetaceans. *Proc. R. Soc. Lond. B* **277,** 3097–3104 (2010).
21. Gingerich, P. D. Evolution and the fossil record: patterns, rates, and processes. *Can. J. Zool.* **65,** 1053–1060 (1987).
22. Hunt, G. Fitting and comparing models of phyletic evolution: random walks and beyond. *Paleobiology* **32,** 578–601 (2006).
23. Polly, P. D. Paleontology and the comparative method: ancestral node reconstructions versus observed node values. *Am. Nat.* **157,** 596–609 (2001).
24. Ruta, M., Wagner, P. J. & Coates, M. I. Evolutionary patterns in early tetrapods. I. Rapid initial diversification followed by decrease in rates of character change. *Proc. R. Soc. Lond. B* **273,** 2107–2111 (2006).
25. Venditti, C., Meade, A. & Pagel, M. Phylogenies reveal new interpretation of speciation and the Red Queen. *Nature* **463,** 349–352 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by a Leverhulme Trust Early Career Fellowship (ECF/2009/0029) to C.V., and by grants to M.P. from the Natural Environment Research Council, UK, the Leverhulme Trust and the European Research Council. We thank R. Freckleton for discussion regarding the implementation of our variable-rates model.

Author Contributions C.V., A.M. and M.P. contributed to all aspects of this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.V. (c.venditti@hull.ac.uk) or M.P. (m.pagel@reading.ac.uk).

doi:10.1038/nature10521

Perception of sniff phase in mouse olfaction

Matthew Smear[1,2], Roman Shusterman[1], Rodney O'Connor[1]†, Thomas Bozza[1,2] & Dmitry Rinberg[1]

Olfactory systems encode odours by which neurons respond and by when they respond[1–3]. In mammals, every sniff evokes a precise, odour-specific sequence of activity across olfactory neurons[4–6]. Likewise, in a variety of neural systems, ranging from sensory periphery[7,8] to cognitive centres[9], neuronal activity is timed relative to sampling behaviour and/or internally generated oscillations. As in these neural systems, relative timing of activity may represent information in the olfactory system[10,11]. However, there is no evidence that mammalian olfactory systems read such cues[12,13]. To test whether mice perceive the timing of olfactory activation relative to the sniff cycle ('sniff phase'), we used optogenetics in gene-targeted mice to generate spatially constant, temporally controllable olfactory input. Here we show that mice can behaviourally report the sniff phase of optogenetically driven activation of olfactory sensory neurons. Furthermore, mice can discriminate between light-evoked inputs that are shifted in the sniff cycle by as little as 10 milliseconds, which is similar to the temporal precision of olfactory bulb odour responses[14,15]. Electrophysiological recordings in the olfactory bulb of awake mice show that individual cells encode the timing of photoactivation in relation to the sniff in both the timing and the amplitude of their responses. Our work provides evidence that the mammalian olfactory system can read temporal patterns, and suggests that timing of activity relative to sampling behaviour is a potent cue that may enable accurate olfactory percepts to form quickly[11,16].

If mice perceive the timing of olfactory activation, they should be able to discriminate between identical sensory stimuli presented at different times in the sniff cycle. In order to isolate this cue, we used optogenetics[17] to deliver spatially fixed, temporally controllable patterns of olfactory sensory neuron (OSN) stimulation. We engineered a mouse line in which all OSNs express channelrhodopsin-2 fused to the yellow fluorescent protein (ChR2–YFP) from the Olfactory Marker Protein (OMP) locus (Fig. 1a). In OMP–ChR2 mice, ChR2–YFP is expressed in all mature olfactory sensory neurons and their nerve terminals in glomeruli of the olfactory bulb (Fig. 1b).

To establish that we can stimulate the olfactory system with light in these mice, we first tested light detection in OMP–ChR2 ($n = 12$) and wild-type littermate controls ($n = 4$). We implanted a pressure cannula into one nasal cavity to measure sniffing, and an optical fibre in the contralateral cavity for photostimulation (Fig. 1c; see Methods). We tested these mice in a head-fixed, go/no-go task in which mice report perceptual judgments by licking or not (Fig. 1d; see Methods). We first trained mice to report odour detection. All mice achieved above-chance behavioural performance in their first session (binomial test, $P < 0.01$, 200–400 trials), and performed >90% in subsequent sessions (Fig. 1e, Supplementary Fig. 1). After at least four odour sessions, we replaced odour stimuli with light pulses (5 mW power, 1 ms duration). Under these conditions, all OMP–ChR2 mice reported detection of light with similar accuracy as for odour, within the first session (Fig. 1e), while all wild-type mice failed to report light detection above chance level in any of four sessions (binomial test, $P > 0.05$). This shows that light drives behaviour through ChR2-mediated OSN activation.

To test whether the animals perceive the sniff phase of OSN activation, we trained mice ($n = 8$) to discriminate between light stimuli solely on the basis of this cue. In each trial, a single light stimulus occurred, and across trials, stimulus intensity and duration were held constant. Stimuli were delivered 32 ms after the onset of inhalation ('go' sniff phase) or 32 ms after the onset of exhalation ('no-go' sniff phase; Fig. 2a, Supplementary Fig. 2a; see Methods). After switching

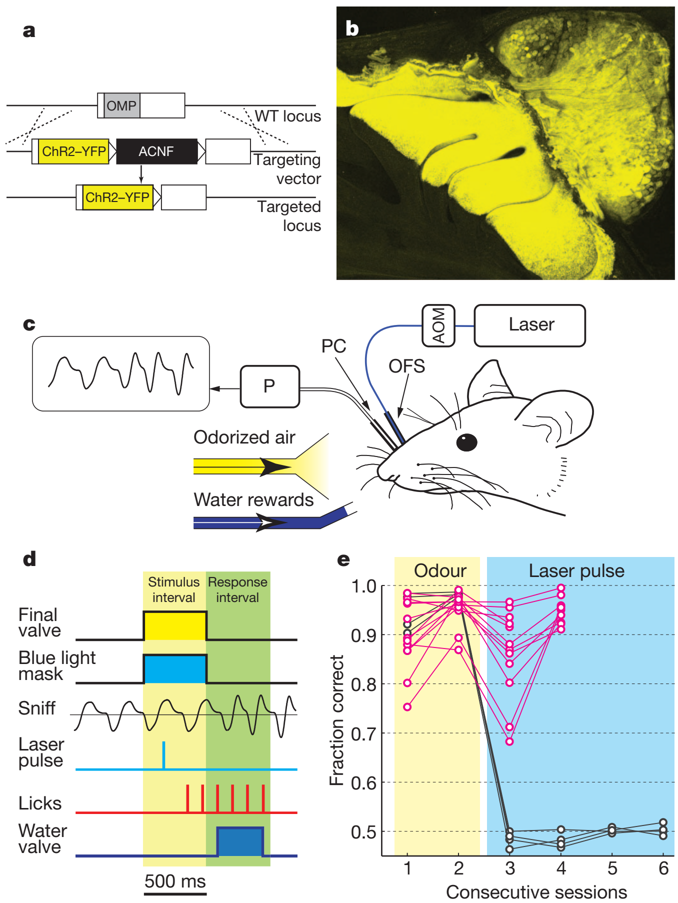


Figure 1 | Stimulating olfaction with light. a, Diagram of the gene targeting strategy. The ChR2–YFP sequence (yellow box) replaces that of OMP (grey box). The targeting selection cassette (ACNF) was removed by germline excision, leaving behind a single loxP site (triangle). **b**, Sagittal view of whole-mount olfactory epithelium and bulb. In OMP–ChR2 mice, ChR2–YFP labels OSNs and their axons in the bulb. **c**, Schematic of experimental set-up. Mice were implanted with a nasal optical fibre stub (OFS) to deliver light, gated by an acousto-optic modulator (AOM). A nasal pressure cannula (PC) coupled to a pressure sensor (P) measures sniffing. Inverted intranasal pressure signal is shown at top left. **d**, Behavioural trial structure. Each trial comprises a stimulus interval (yellow shading) and a response interval (green). **e**, Performance of OMP–ChR2/+ mice (pink circles; $n = 12$) and +/+ littermate controls (black circles; $n = 4$) in odour detection sessions (yellow shading), followed by light detection sessions (blue shading).

[1]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia 20147, USA. [2]Department of Neurobiology, Northwestern University, Evanston, Illinois 60208, USA. †Present address: Department of Biology, Boston University, Boston, Massachusetts 02215, USA.

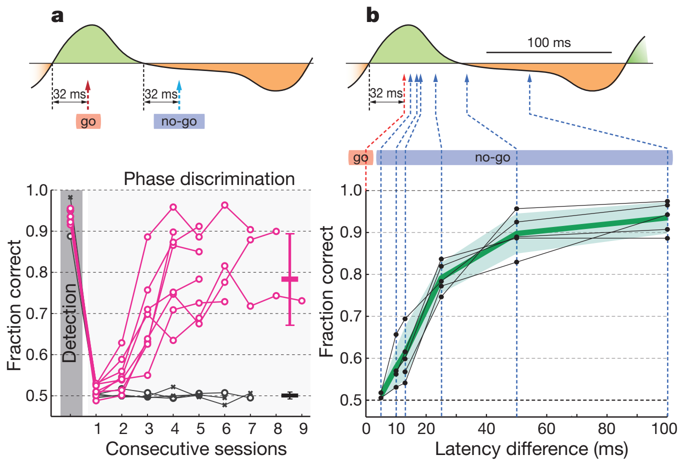


Figure 2 | OMP–ChR2 mice perceive sniff phase. **a**, Top, schematic of the sniff phase discrimination problem, shown relative to a typical sniff waveform, with inhalation shaded green and exhalation shaded orange. Light was delivered 32 ms after inhalation onset (red arrow) in 'go' trials or 32 ms after exhalation onset (blue arrow) in 'no-go' trials. Bottom, connected pink circles show the performance of OMP–ChR2/+ mice ($n = 8$) for their last light detection session, followed by phase discrimination sessions. Black lines show click detection (grey shading) and click phase discrimination performance for OMP–ChR2/+ (black circles, $n = 2$) and wild-type (black asterisks, $n = 2$). Horizontal dashes at right show mean ± s.d. phase discrimination performance for light (pink; $n = 30$ sessions) and click (black; $n = 19$ sessions) stimuli, from the third session onward. **b**, Performance as a function of latency difference. Top, as **a**. Bottom, black filled circles show performance of individual OMP–ChR2 mice while green line and shaded region give mean ± s.d. ($n = 5$).

from detection to the phase discrimination task, mice learned quickly, attaining above-chance performance in their second or third session (Fig. 2a). This behavioural performance demonstrates that mice perceive the sniff phase of olfactory input. In contrast, another set of mice ($n = 4$), tested with an easily detectable auditory click stimulus, failed to report the sniff phase of clicks (Fig. 2a, binomial test, $P > 0.05$). This suggests that the olfactory system may have unique access to sniff timing information.

How acute is the mouse's sense of time in the sniff? To test whether mice can discern finer timing differences, we trained five mice to discriminate between light stimuli occurring at the same 'go' sniff phase as above, and those occurring with varying latencies (5–100 ms) after the 'go' sniff phase (Fig. 2b, top). A single 'no-go' latency was tested in each session. Performance is high for 'no-go' latencies of 50 ms or greater (Fig. 2b). Mice maintained high accuracy at 25 ms latency (Supplementary Fig. 2b; 80 ± 5%, mean ± s.d.), and four of five mice exceeded chance performance at 10 ms (binomial test, $P < 0.01$). Achieving this performance does not require extensive training—at each latency, all mice performed three or fewer sessions. Across sessions, sniff durations do not differ systematically, but do vary from trial to trial, and mice performed better in trials with short inhalation duration (Supplementary Fig. 3). These data show that mice can sense timing differences that are tenfold shorter than a sniff cycle.

To characterize how the olfactory bulb responds to the stimuli presented in our behavioural experiments, we recorded light-evoked responses from 86 neurons, putatively mitral/tufted (M/T) cells, in the olfactory bulb of five OMP–ChR2 mice (see Methods). These mice were awake but were not performing a task. Out of 86 cells, 57% exhibited light-evoked responses: 26 gave excitatory responses, while 23 cells gave inhibitory responses (see Methods). By comparison, in a recent study, it was found that individual odours, on average, evoke responses in 66% of M/T cells[15]. Therefore, the light stimulus used in our behavioural experiments activates a similar number of M/T cells as do odours.

We then delivered light stimuli at a range of latencies relative to sniff (2–6 latencies per recording session). Some cells responded strongly at all latencies tested (for example, Fig. 3a, cell 1). In contrast, other cells exhibited varying response amplitude with stimulus latency (for example, Fig. 3a, cell 2).

To quantify the temporal dynamics of excitatory light responses, we fitted a Gaussian function to the difference between inhalation-aligned spike histograms with and without stimuli. The fit parameters yield measures of latency (τ), duration (σ) and amplitude (A) (see Methods). The brief durations (Fig. 3b) and narrow latency distribution (Fig. 3c) of these responses demonstrate that M/T cells faithfully propagate the timing of OSN photostimulation to their central targets. In addition, cells may vary their response amplitudes when stimulated at different times in the sniff cycle (for example, Fig. 3a, cell 2). Tuning curves for latency relative to sniff were heterogeneous across cells and often non-monotonic (Fig. 3c; tuning curves for cells receiving stimuli at six latencies are shown in Supplementary Fig. 4). As a result of this tuning, information about timing of OSN activation is also contained in the pattern of response amplitudes across the M/T cell ensemble. Consequently, olfactory bulb responses contain two cues that may enable the animal to report the latency of light stimuli relative to sniff onset: timing and amplitude.

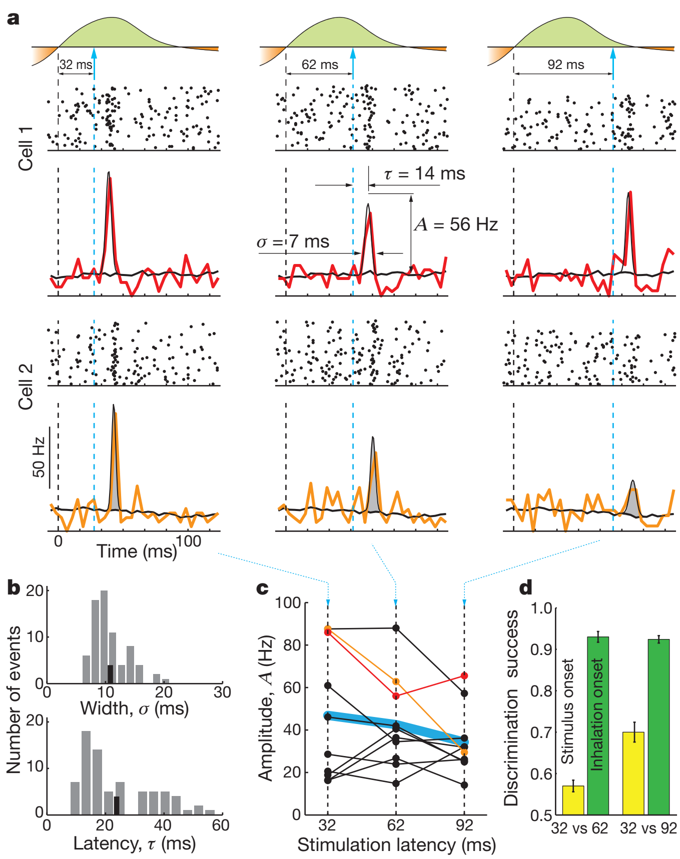


Figure 3 | Response of mitral/tufted cells to light stimulation. **a**, Top, light application. Middle and bottom, raster plots + PSTH for two cells' responses (cell 1, top; cell 2, bottom) to light at three latencies (32 ms, left; 62 ms middle; 92 ms, right) after inhalation onset. Coloured lines are PSTHs for light responses. Thick grey lines are PSTHs for spontaneous activity. Thin black line is a Gaussian fit of the difference between PSTHs for stimulated and unstimulated sniffs. The fit parameters yield measures of response width (σ), latency (τ) and amplitude (A). **b**, Distribution of response widths (σ, top) and latencies (τ, bottom). Grey bars, data; thick black line, mean. **c**, Connected filled circles show response amplitudes (A) from individual cells (red and orange dots show respectively cells 1 and 2). Blue line indicates the across-cell mean. **d**, Classification performance for the neuronal population response (mean ± s.d. across repeated permutations) discriminating between 32 and 62 ms and between 32 and 92 ms light stimulation latency. Responses were aligned to the stimulus onset (yellow) and inhalation onset (green).

We estimated the amount of information that neuronal response timing and amplitude carry about stimulus sniff-timing using a classification algorithm. In essence, this algorithm measures how well stimulus sniff phases can be discriminated on the basis of each trial's neuronal response cues (51 cells; see Methods). We compared classification success with and without the temporal cue, by applying the classification algorithm to the same neural data aligned in two ways: to inhalation onset (as in Fig. 3a), in which case both the sniff-timing and amplitude cues are available, or to stimulus onset, which eliminates the timing cue. When the classification algorithm is applied to sniff-aligned data, discrimination success for both pairs of stimuli is above 90% (Fig. 3d), comparable to behavioural performance (Fig. 2b). When applied to stimulus-aligned data, the classification algorithm performs worse, yet still above chance for both pairs of stimuli (Fig. 3d; 32 versus 62: 57 ± 1.4%; 32 versus 92: 70 ± 2.4%; see Methods). These analyses indicate that both timing and amplitude cues carry information about stimulus latency relative to sniff.

Our work provides evidence (the first, to our knowledge) that the mammalian olfactory system can read temporal information and reveals the striking temporal acuity of this mechanism. Mice discriminate between inputs solely on the basis of 'sniff phase'—that is, relative time in the sniff cycle. Whether this temporal cue is read in phase or time coordinates remains an open question[14,15]. Although mice can use timing relative to sniff in our experimental design, where this is the only cue available, it is unclear to what degree animals do so in the context of natural odour stimulation. However, the ease with which mice sense the sniff phase of olfactory input in isolation argues that this cue plays an important role in representing odours.

What mechanisms temporally pattern olfactory responses? Recent work in insects has revealed odour-specific temporal structure in OSN responses[18,19], which can largely be explained by a simple kinetic model of ligand-receptor binding[19]. Rodent OSNs also give temporally structured responses: calcium-imaging studies demonstrate odour-specific sniff-locked sequences across glomeruli[6]. These sniff-locked sequences may propagate to recipient neurons in the olfactory bulb, where additional sniff-entrained circuit mechanisms may operate. For example, M/T cells exhibit sniff-locked oscillations of membrane potential[16,20], which can transform input intensity into timing by modulating excitability[10,21]. Consistent with this idea, raising odour concentration shifts activation to earlier times in the sniff cycle[16,20]. The resultant mapping of intensity to time agrees with a theory in which latency relative to the sniff onset encodes the intensity of receptor activation[10,16,22]. Determining whether these or other mechanisms account for how the olfactory system transforms odours to sniff-locked temporal patterns will require further investigation. We provide a strong impetus for such work by showing that mice can perceive sniff phase cues.

The behavioural relevance of temporal coding has received experimental support in a variety of sensory systems[23–25]. Despite the long history of work on the temporal patterning of odour responses[4–6], the behavioural relevance of temporal coding in olfaction has not been demonstrated, despite a previous attempt to do so[12]. Our strategy of timing optogenetic activation relative to a putative timing reference signal, which has proven successful for olfaction, could be generalized to test phase/latency coding hypotheses in other systems[26]. For example, timing relative to sampling behaviour and/or local field potential oscillations has been proposed as a coding variable in vibrissal somatosensation[27] and vision[8,28]. Beyond more peripheral sensory areas, our optogenetic strategy may also be applicable to putative temporal cues in more central systems—for example, the phase precession of hippocampal place cells[9].

The ability of mice to discriminate between identical patterns of illuminated OSNs solely on the basis of timing with respect to the sniff suggests that differences in the spatial pattern of OSN activation may be unnecessary for perceptible olfactory differences, contrary to prior suggestions[13]. However, stimulus- and sniff-related signals may converge within OSNs, creating sniff-phase-dependent spatial patterns of OSN activation. In some OSNs, activity is modulated by the sniff cycle in the absence of overt odour stimulation, perhaps responding to airflow or background odour[29]. Downstream of OSNs, stimulus- and sniff-related signals may be integrated by olfactory bulb circuits, or further downstream in the olfactory system. These considerations underscore the fact that spatial coding and temporal coding are not mutually exclusive, and may instead exhibit synergy in numerous ways. We speculate that time comparisons across glomeruli give a concentration-invariant readout for odour identity[11,16,22], whereas temporal comparison to an internal representation of the sniff yields information about odour concentration. Such a coding scheme can rapidly resolve ambiguities that arise as odour identity and intensity change[11]. Extracting both parameters on a sniff-by-sniff basis may help animals locate and identify odour sources in natural olfactory scenes.

METHODS SUMMARY

OMP–ChR2 heterozygous mice and wild-type mice were implanted with headbars for head fixation, pressure cannulae in the nasal cavity for sniff recording and optic fibre stubs in the contralateral nasal cavity for light stimulus delivery. After at least three days of recovery followed by at least ten days of restriction of water intake to 1 ml d^{-1}, mice were trained to lick for water while head-fixed in a behavioural box. Then, mice were trained to perform the following behavioural tasks in a go/no-go paradigm: (1) odour detection; lick in response to odour and do not lick to blank delivery, (2) light detection; lick in response to light stimulation via nasal optic fibre (5 mW power, 1 ms duration), (3) sniff phase discrimination; lick in response to light stimulation at fixed latency after inhalation onset (32 ms) ('go' stimulus) and do not lick in response to light stimulation triggered after exhalation onset ('no-go' stimulus); and (4) fine temporal discrimination: lick in response to light stimulation at fixed latency after inhalation onset ('go' stimulus), and do not lick in response to light stimulation delayed from the 'go' stimulus latency by some time interval (5–100 ms, 'no-go' stimulus). In each session (~400 trials), one 'no-go' stimulus latency was used. Onsets of inhalation and exhalation were defined as zero-crossings of the intranasal pressure signal. For electrophysiological recordings, mice were also implanted with 16- or 32-channel silicon probes. In recording sessions, mice were awake but not performing a task. Light stimuli were triggered at fixed latencies (32, 62 and 92 ms) after onsets of inhalation or exhalation phase.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 April; accepted 2 September 2011.
Published online 12 October 2011.

1. Laurent, G. Olfactory network dynamics and the coding of multidimensional signals. *Nature Rev. Neurosci.* **3,** 884–895 (2002).
2. Friedrich, R. W. & Laurent, G. Dynamic optimization of odor representations by slow temporal patterning of mitral cell activity. *Science* **291,** 889–894 (2001).
3. Junek, S., Kludt, E., Wolf, F. & Schild, D. Olfactory coding with patterns of response latencies. *Neuron* **67,** 872–884 (2011).
4. Macrides, F. & Chorover, S. L. Olfactory bulb units: activity correlated with inhalation cycles and odor quality. *Science* **175,** 84–87 (1972).
5. Chaput, M. & Holley, A. Single unit responses of olfactory bulb neurones to odour presentation in awake rabbits. *J. Physiol. (Paris)* **76,** 551–558 (1980).
6. Spors, H., Wachowiak, M., Cohen, L. B. & Friedrich, R. W. Temporal dynamics and latency patterns of receptor neuron input to the olfactory bulb. *J. Neurosci.* **26,** 1247–1259 (2006).
7. Szabo, T. & Hagiwara, S. A latency-change mechanism involved in sensory coding of electric fish (mormyrids). *Physiol. Behav.* **2,** 331–335 (1967).
8. Gollisch, T. & Meister, M. Rapid neural coding in the retina with relative spike latencies. *Science* **319,** 1108–1111 (2008).
9. O'Keefe, J. & Recce, M. L. Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3,** 317–330 (1993).
10. Hopfield, J. J. Pattern recognition computation using action potential timing for stimulus representation. *Nature* **376,** 33–36 (1995).
11. Schaefer, A. T. & Margrie, T. W. Spatiotemporal representations in the olfactory system. *Trends Neurosci.* **30,** 92–100 (2007).
12. Monod, B., Mouly, A. M., Vigouroux, M. & Holley, A. An investigation of some temporal aspects of olfactory coding with the model of multi-site electrical stimulation of the olfactory bulb in the rat. *Behav. Brain Res.* **33,** 51–63 (1989).
13. Leon, M. & Johnson, B. A. Is there a spacetime continuum in olfaction? *Cell. Mol. Life Sci.* **66,** 2135–2150 (2009).
14. Cury, K. M. & Uchida, N. Robust odor coding via inhalation-coupled transient activity in the mammalian olfactory bulb. *Neuron* **68,** 570–585 (2010).
15. Shusterman, R., Smear, M., Koulakov, A. & Rinberg, D. Precise olfactory responses tile the sniff cycle. *Nature Neurosci.* **14,** 1039–1044 (2011).

16. Margrie, T. W. & Schaefer, A. T. Theta oscillation coupled spike latencies yield computational vigour in a mammalian sensory system. *J. Physiol. (Lond.)* **546,** 363–374 (2003).
17. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neurosci.* **8,** 1263–1268 (2005).
18. Raman, B., Joseph, J., Tang, J. & Stopfer, M. Temporally diverse firing patterns in olfactory receptor neurons underlie spatiotemporal neural codes for odors. *J. Neurosci.* **30,** 1994–2006 (2010).
19. Nagel, K. I. & Wilson, R. I. Biophysical mechanisms underlying olfactory receptor neuron dynamics. *Nature Neurosci.* **14,** 208–216 (2011).
20. Cang, J. & Isaacson, J. S. In vivo whole-cell recording of odor-evoked synaptic transmission in the rat olfactory bulb. *J. Neurosci.* **23,** 4108–4116 (2003).
21. Perkel, D. H. & Bullock, T. H. Neural coding. *Neurosci. Res. Prog. Bull.* **6,** 219–349 (1968).
22. Brody, C. D. & Hopfield, J. J. Simple networks for spike-timing-based computation, with application to olfactory processing. *Neuron* **37,** 843–852 (2003).
23. Hall, C., Bell, C. & Zelick, R. Behavioral evidence of a latency code for stimulus intensity in mormyrid electric fish. *J. Comp. Physiol. A* **177,** 29–39 (1995).
24. Di Lorenzo, P. M., Leshchinskiy, S., Moroney, D. N. & Ozdoba, J. M. Making time count: functional evidence for temporal coding of taste sensation. *Behav. Neurosci.* **123,** 14–25 (2009).
25. Jacobs, A. L., Fridman, G., Douglas, R. M., Alam, N. M. & Latham, P. Ruling out and ruling in neural codes. *Proc. Natl Acad. Sci. USA* **106,** 5936–5941 (2009).
26. VanRullen, R., Guyonneau, R. & Thorpe, S. J. Spike times make sense. *Trends Neurosci.* **28,** 1–4 (2005).
27. Curtis, J. C. & Kleinfeld, D. Phase-to-rate transformations encode touch in cortical neurons of a scanning sensorimotor system. *Nature Neurosci.* **12,** 492–501 (2009).
28. Montemurro, M. A., Rasch, M. J., Murayama, Y., Logothetis, N. K. & Panzeri, S. Phase-of-firing coding of natural visual stimuli in primary visual cortex. *Curr. Biol.* **18,** 375–380 (2008).
29. Grosmaitre, X., Santarelli, L. C., Tan, J., Luo, M. & Ma, M. Dual functions of mammalian olfactory sensory neurons as odor detectors and mechanical sensors. *Nature Neurosci.* **10,** 348–354 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank L. Doglio and the Transgenic and Targeted Mutagenesis Laboratory at Northwestern University for generation of chimaeric mice, B. Weiland for technical help with cloning and gene targeting, D. Huber, D. O'Connor and T. Komiyama for advice on mouse behaviour, D. Wesson and M. Wachowiak for instruction on sniff measurement, J. Nunez-Iglesias for assistance with statistics, G. Shtengel for advice on laser set-up, and T. Tabachnik and H. Davidowitz for help designing the behavioural rig. J. Osborne fabricated the microdrive. G. Lott provided digital acquisition software. A. Koulakov contributed to spike-sorting and classification algorithms. We thank W. Denk, K. Svoboda, R. Gütig, R. Egnor, M. Orger and A. Resulaj for comments on the manuscript. This work was supported by the Visiting Scientist Program at JFRC. T.B. was supported by NIDCD (R01DC009640, R21DC010911), the Whitehall Foundation and the Brain Research Foundation.

Author Contributions M.S. and D.R. designed the study and build the experimental set-up, M.S. performed the experiments and analysed the behavioural data. R.S. and M.S. performed the electrophysiological recordings, R.S. and D.R. analysed the electrophysiological data, and T.B. initiated the transgenic approach and generated the gene-targeted mice. R.O. developed the laser optics and optical fibre design. M.S., T.B. and D.R wrote the manuscript. D.R. and T.B. supervised the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.R. (rinbergd@janelia.hhmi.org) or T.B. (bozza@northwestern.edu).

METHODS

Gene targeting. The coding sequence for ChR2(H134R)–YFP (gift of G. Nagel, Max Planck Institute for Biophysics) was amplified and cloned into an OMP targeting vector[30] replacing the endogenous OMP coding sequence. The vector also contained an autoexcising *neo* selection cassette[31]. The vector was linearized and electroporated into E14 ES cells, and correctly targeted clones isolated using standard methods. Targeted clones were injected into C57BL/6J blastocysts to generate chimaeras. The allele was passed through the male germline, removing the *neo* cassette. The OMP–ChR2–YFP line was derived from clone 'OCY-58'. This strain will be made available through The Jackson Laboratory (Tyr<c-2J> -OMP<tm1(COP4/EYFP)-Tboz>/J; STOCK #14173); address requests for information to T.B.

Animals. Data were collected in 15 OMP–ChR2–YFP heterozygous mice and 4 wild-type littermates. All mice had at least one normal copy of OMP. Subjects were 6–8 weeks old at the beginning of behavioural training and were maintained on a 12 h light/dark cycle (lights on at 8:00 p.m.) in isolated cages in a temperature- and humidity-controlled animal facility. All animal care and experimental procedures were in strict accordance with a protocol approved by the Howard Hughes Medical Institute Institutional Animal Care and Use Committee.

Sniff recording. To monitor the sniff signal, a 7-mm-long stainless cannula (gauge 23, Small Parts capillary tubing) was implanted in the nasal cavity. The cannula was capped between experimental recordings. During experiments, the cannula was connected via polyethylene tubing (801000, A-M systems) to a pressure sensor (MPX5050, Freescale Semiconductor) and custom-made preamplifier circuit. The signal from the preamplifier was amplified 20× and low-pass-filtered below 20 Hz (Cygnus Technology), digitized with an NIDAQ board (National Instruments) and acquired by an in-house data acquisition program (SpikeHound, written by G. Lott). The pressure signal was also sent to a custom-made comparator board that created a square TTL pulse between rising and falling zero crossings in the pressure signal. This pulse went to a behavioural control board to trigger light stimuli (see below).

Surgery. Mice were anaesthetized using isofluorane gas anaesthesia. The horizontal bar for head fixation, pressure cannula, optic fibre stub, and, in a subset of mice, electrode chamber were implanted during a single session of surgery. To implant the sniffing cannula, a small hole was drilled in the bone overlying the nasal cavity, into which the cannula was inserted and affixed with glue and stabilized with dental cement. The optic fibre stub was implanted and fixed in the same way in the contralateral nasal cavity. To implant the electrode chamber, a small craniotomy (~1 mm^2) was opened above the olfactory bulb, roughly centred along the A–P and M–L axes of the bulb. An electrode chamber with a silicon probe was fixed by dental cement to the skull, posterior to the olfactory bulb. The reference electrode was implanted in the cerebellum. After surgery, a mouse was caged individually and given at least 3 days for recovery.

Stimulus delivery. For odour stimulus delivery, we used a nine-odour air dilution olfactometer. Odorants (Sigma-Aldrich) were stored in liquid phase in dark vials. The airflow through the selected odorant vial was diluted 10 times by the main airflow stream and homogenized in a long thin capillary before reaching the final valve. Between stimuli, a steady stream of 1,000 ml min^{-1} of clean air flowed to the odour port continuously, while the flow from the olfactometer was directed to an exhaust. During stimulus delivery, the final valve (four-way Teflon valve; NResearch) switched the odour flow to the odour port, and diverted the clean airflow to the exhaust. Temporal odour concentration profile was checked by a mini photoionization detector (miniPID, Aurora Scientific). The concentration reached a steady state 25–40 ms after final valve opening.

A 473-nm laser (Ciel Blue DPSS, Photonic Solutions) was our light source. The main beam was split to provide stimulus for two experiments. Each secondary beam was gated by an acousto-optic modulator (AOM, QuantaTech), which enabled analogue control of light stimulus power with microsecond timing precision. A fibre launcher (Thorlabs) was positioned to catch the first mode from the AOM in a 100 μm core multi-mode optic fibre. The amplitude of the square pulse controlled the angle of AOM beam diversion, thus providing fine control of power collected by the fibre launcher.

The opposite end of the fibre terminated in a ceramic ferrule (Precision Fibre Products), which could be coupled via an phosphor-bronze sleeve (Optequip) to an identical ceramic ferrule holding the optical fibre stub implanted into the mouse. The light stimulus power at the ferrule ending was measured by a powermeter (Thorlabs), and calibrated daily by adjusting the amplitude of the pulse to the AOM driver. The ferrule coupling allowed efficient transmission of light (80–90%), but also leaked light. To prevent the mouse from using this leaked light as a visual cue, two bright blue LEDs (Luxeon V-star, Philips Lumileds Lighting Company) were positioned on either side of the mouse's head, about 1 cm from each eye. These LEDs were activated during the stimulus period of each trial, to mask leak light from the laser.

Water delivery was based on gravitational flow controlled by a solenoid valve (Clippard) connected via Tygon tubing to a stainless steel cannula (gauge 21, Small Parts), which served as a lick tube. The lick tube was positioned near the animal's mouth, and could be moved by a micromanipulator. The water volume was controlled by the duration of valve opening for 200–400 ms duration, calibrated daily to give approximately 5 μl per opening. Licks were detected by photodiode beam break by the mouse's tongue.

Behavioural control. All behavioural events (odour and final valve opening, laser delivery, water delivery, and photobeam crossing) were monitored and controlled by a behavioural board (LASOM1, RPMetrix), which allowed real-time experimental control with millisecond precision. The behavioural board reads trial parameters and sends trial results to a PC running custom-written MatLab routines (Mathworks).

Behavioural task and training. After at least 3 days of post-operative recovery and at least 7 days of water restriction (1 ml d^{-1}), we began to train the mice. Training started with water-sampling sessions, in which the mouse was placed in the head fixation set-up and given water for licking. Before moving to the next stage of training, each mouse had to perform two sessions in which it licked enough to receive its full 1 ml of water for the day. Mice that failed to collect their full daily ration in a behavioural session were supplemented with water in their home cage.

Next, the mice were trained to report odour detection. A behavioural session was broken into pseudo-randomly ordered trials, each of which consisted of a stimulus period, a response period, and an intertrial interval (ITI). During the stimulus period, the final valve switched to direct air from the olfactometer to the animal. Olfactometer flow passed through a vial containing liquid odorant, or through a blank vial. Mice received water for licks during the response period following odour delivery, and did not receive water if they licked in response to blank delivery. These incorrect licks were punished by lengthened ITI. Correct trial ITIs were 5,000 ms plus a random number between 1 and 2,000 ms, while incorrect trial ITIs were 10,000 ms plus a random number between 1 and 6,000 ms. Randomization of ITIs was intended to prevent the possibility that animals would anticipate stimulus delivery and synchronize their sniffing.

After at least two sessions with a first odorant, another odour became the 'go' stimulus. In pilot experiments, we found that mice usually would not lick for a new stimulus under these conditions, but only for the initially trained odorant. In order to facilitate more rapid acquisition of licking for new stimuli, we included an 'associative block' at the beginning of every session. The associative block consisted of ten water valve openings delivered immediately after fixing the mouse in the behavioural chamber, followed by 20 consecutive 'go' trials with the new stimulus. Acquisition blocks were included in all behavioural sessions reported here.

Light sessions began after at least four odour detection sessions for each mouse. Licking in response to light stimulus was rewarded with water. Licking when no light stimulus was delivered lengthened the ITI. For all light sessions, the stimulus power at the ferrule coupling was 5 mW, while the stimulus duration was 1 ms. Pilot experiments suggested that stimuli roughly an order of magnitude more powerful or longer duration could be detected by +/+ mice. All OMP–ChR2/+ mice reported light detection. Only those mice that maintained a good sniff signal could be tested for the temporal discriminations.

After at least one session of light detection, mice began sniff phase discrimination sessions. In these, light was triggered from a rising or falling zero crossing in the sign-inverted pressure signal, which indicate the onset of inhalation or exhalation, respectively. Reliable detection of zero-crossing events was facilitated by low pass filtering sniff signals below 20 Hz, which introduced a constant delay of 32 ms, as described in the text. After at least three sniff phase discrimination sessions, those mice that maintained a good sniff signal were tested for the finer latency discriminations. In each of these sessions, a single no-go stimulus was used, and each mouse did three or fewer sessions at each latency, in descending order.

Electrophysiology. Mitral/tufted cell spiking activity was recorded using 16- or 32-channel silicon probes (NeuroNexus, models a2x2-tet-3mm-150-150-312(F16), a4x8-5mm 150-200-312(F32)). Cells were recorded in the mitral cell layer of the dorsal bulb, 300–400 μm from the bulb surface. The identity of M/T cells were established based on the criteria formulated in previous work[32]. The data were acquired using 32-channel data acquisition system (Digital Lynx, NeuraLynx) with widely open broadband filters and sampling frequency 0.1–9,000 Hz.

Data analysis and spike extraction. Data analysis was done in MatLab (Mathworks). Acquired electrophysiological data were filtered and spike sorted. For Si-probe data we used the M-Clust program (written by A. D. Redish) and a software package (written by A. Koulakov).

Light responses. To identify excitatory and inhibitory responses in neurons, we used a randomization test to compare the distributions of total spike counts of each cell with and without light stimulation[33]. In each session, 3,000–5,000 sniff cycles without light stimulation nor following light stimulation within 5 cycles were

defined as control cycles. N cycles with light stimulation (one sniff cycle per trial with stimulus) were test cycles ($N=22$–70). For time windows of 1, 2, ... 100 ms after stimulus onset, we counted the number of spikes in the time window across the N trials. We then compared this number against the distribution of spike counts in the same time window for randomly chosen subsets of size N from the control cycles. The P value was estimated as the proportion of control spike counts larger than the observed test count, relative to the distribution median, multiplied by 2 to account for the two-sidedness of the test. We considered a cell to respond to the stimulus if the P value for at least one time window was less than 0.003, which corresponded to a false discovery rate of 0.05 by the Benjamini-Hochberg procedure. For statistically significant excitatory responses, we fitted a Gaussian, $f = A\exp[-\pi(t-\tau)^2/\sigma^2]$, to the difference between spike histograms for stimulated and unstimulated sniffs, where A is the amplitude of the response, τ is its average latency, and σ is its average width. The parameter σ is chosen so that $A \times \sigma$ is equal to the integral below the Gaussian function and corresponds to the average number of extra spikes per trial in response to the stimulus.

Classification analysis. To estimate how well a population of neurons ($n = 51$) can discriminate between two stimuli on a single trial, we used a template-matching algorithm [15]. For each pair of stimulus latencies (light stimulation at 32 and 62 ms latency, and at 32 and 92 ms latency), we aligned neuronal signals in two ways, relative to the inhalation onset or relative to the stimulus onset. For every trial we built a response vector $\mathrm{r}_k = \{r_{1,1}, r_{1,2}, \ldots r_{1,m}, r_{2,1}, \ldots r_{2,m}, \ldots r_{n,m}\}$, where individual components, $r_{i,j}$, were number of spikes in a time bin $j(j=1,\ldots m)$ of a neuron $i(i=1,\ldots n)$. Time bins of 10 ms covered the interval from either the onset of inhalation or stimulus for the duration of 150 ms. For every stimulus and every trial, we estimated template vectors $\mathrm{r}_{k,s} = \langle r_i \rangle_{i, i\neq k, S(i)=s}$ (where $S(i)$ is a stimulus type for a trial i, and $\langle\rangle_i$ is averaging over i), which averaged all trials for each stimuli $s = 1{,}2$, excluding the given trial k. Then we assigned a given trial to one of the templates based on the shortest Euclidian distance between the response vector $\mathbf{r}_k$ and the templates' vectors $\bar{\mathrm{r}}_{k,s}$. The classification success was equal to a portion of correct assignments.

30. Bozza, T., McGann, J. P., Mombaerts, P. & Wachowiak, M. In vivo imaging of neuronal activity by targeted expression of a genetically encoded probe in the mouse. *Neuron* **42,** 9–21 (2004).
31. Bunting, M., Bernstein, K. E., Greer, J. M., Capecchi, M. R. & Thomas, K. R. Targeting genes for self-excision in the germ line. *Genes Dev.* **13,** 1524–1528 (1999).
32. Rinberg, D., Koulakov, A. & Gelperin, A. Sparse odor coding in awake behaving mice. *J. Neurosci.* **26,** 8857–8865 (2006).
33. Garthwaite, P. H., Jolliffe, I. T. & Jones, B. *Statistical Inference* (Oxford Univ. Press, 2002).

doi:10.1038/nature10517

Killer cell immunoglobulin-like receptor 3DL1-mediated recognition of human leukocyte antigen B

Julian P. Vivian[1], Renee C. Duncan[1], Richard Berry[1], Geraldine M. O'Connor[2], Hugh H. Reid[1], Travis Beddoe[1], Stephanie Gras[1], Philippa M. Saunders[3], Maya A. Olshina[1], Jacqueline M. L. Widjaja[3], Christopher M. Harpur[3], Jie Lin[3], Sebastien M. Maloveste[4], David A. Price[5,6], Bernard A. P. Lafont[4], Daniel W. McVicar[2], Craig S. Clements[1], Andrew G. Brooks[3] & Jamie Rossjohn[1,5]

Members of the killer cell immunoglobulin-like receptor (KIR) family, a large group of polymorphic receptors expressed on natural killer (NK) cells, recognize particular peptide-laden human leukocyte antigen (pHLA) class I molecules and have a pivotal role in innate immune responses[1]. Allelic variation and extensive polymorphism within the three-domain KIR family (KIR3D, domains D0–D1–D2) affects pHLA binding specificity and is linked to the control of viral replication and the treatment outcome of certain haematological malignancies[1–3]. Here we describe the structure of a human KIR3DL1 receptor bound to HLA-B*5701 complexed with a self-peptide. KIR3DL1 clamped around the carboxy-terminal end of the HLA-B*5701 antigen-binding cleft, resulting in two discontinuous footprints on the pHLA. First, the D0 domain, a distinguishing feature of the KIR3D family, extended towards β2-microglobulin and abutted a region of the HLA molecule with limited polymorphism, thereby acting as an 'innate HLA sensor' domain. Second, whereas the D2–HLA-B*5701 interface exhibited a high degree of complementarity, the D1–pHLA-B*5701 contacts were suboptimal and accommodated a degree of sequence variation both within the peptide and the polymorphic region of the HLA molecule. Although the two-domain KIR (KIR2D) and KIR3DL1 docked similarly onto HLA-C[4,5] and HLA-B respectively, the corresponding D1-mediated interactions differed markedly, thereby providing insight into the specificity of KIR3DL1 for discrete HLA-A and HLA-B allotypes. Collectively, in association with extensive mutagenesis studies at the KIR3DL1–pHLA-B*5701 interface, we provide a framework for understanding the intricate interplay between peptide variability, KIR3D and HLA polymorphism in determining the specificity requirements of this essential innate interaction that is conserved across primate species.

HLA-B57 carriage has been associated with delayed progression to AIDS in HIV-infected individuals, with a strong genetic association between the KIR3DL1–HLA-B57 interaction, reduced viral loads and delayed HIV disease progression[3]. We expressed KIR3DL1*001, a prototypical family member, and co-complexed it with HLA-B*5701 bound to a self-peptide (LSSPVTKSF). The affinity (K_D) of this interaction was approximately 17 μM (Supplementary Table 1 and Supplementary Fig. 1). We then determined the KIR3DL1*001–HLA-B*5701–LSSPVTKSF structure to 1.8 Å resolution (Supplementary Table 2 and Supplementary Fig. 2). KIR3DL1*001 clamped around the C-terminal end of the HLA-B*5701 antigen-binding cleft (Fig. 1a, b), forming an extensive interface (total buried surface area (BSA), 1,740 $Å^2$) that encompassed two discontinuous sites—one mediated via the D0 domain and the other via the D1–D2 domains (Figs 1c, d and 2a–d). KIR3DL1*001 adopted an elongated, zigzag conformation, with the three immunoglobulin (Ig) domains, termed D0, D1 and D2 (residues 7–98, 99–198 and 203–292, respectively) defined by the E-type Ig fold topology (Fig. 1a). The D0 domain, a feature of the KIR3D family[6] packed against the D1 domain, the relative juxtapositioning of which (83°) is similar to that of the D1–D2 inter-domain angle (81°), which in turn is analogous to the relative orientation of D1–D2 domains (76°) found in the KIR2D receptors (root mean squared deviation (r.m.s.d.) of D1–D2 domains in KIR2DL1 and

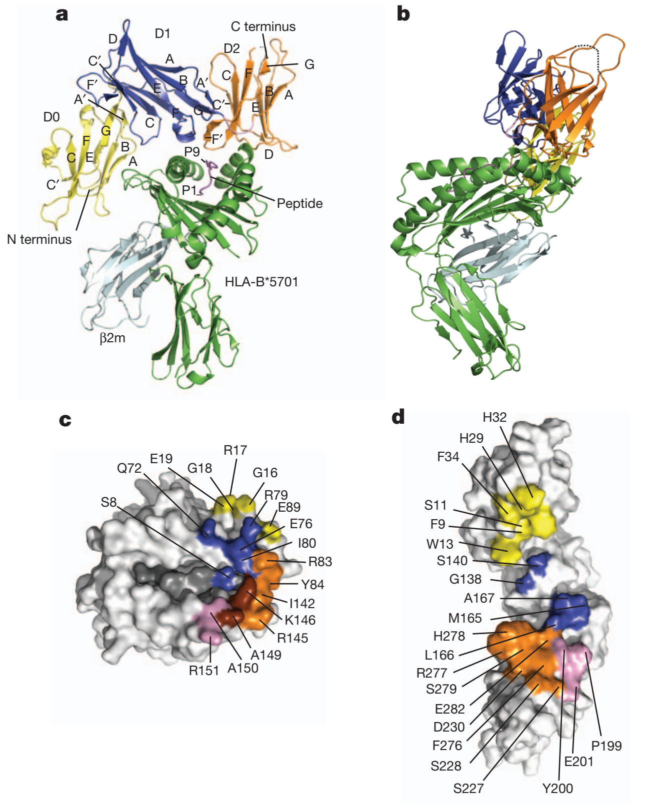


Figure 1 | Structure of the KIR3DL1*001–pHLA-B*5701 complex. **a**, **b**, Orthogonal views of the complex with the KIR3DL1*001 β-strands labelled. The HLA and $β_2$-microglobulin (β2m) are coloured green and cyan, respectively; D0, D1, D1–D2 loop and D2 are coloured yellow, blue, pink and orange, respectively; dashed line represents the unresolved loop between the E and F β-strands. **c**, **d**, The footprint mapped to the surface of HLA and KIR3DL1*001, respectively, with residues coloured in each case according to the interacting KIR3DL1*001 domain: D0 (yellow), D1 (blue), D1–D2 loop (pink) and D2 (orange). Residues that contact the linker and the D2 domain are coloured brown.

[1]Department of Biochemistry and Molecular Biology, School of Biomedical Sciences, Monash University, Clayton, Victoria 3800, Australia. [2]Cancer and Inflammation Program, National Cancer Institute-Frederick, Frederick, Maryland 21702, USA. [3]Department of Microbiology & Immunology, University of Melbourne, Parkville, Victoria 3010, Australia. [4]Non-Human Primate Immunogenetics and Cellular Immunology Unit, Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [5]Department of Infection, Immunity and Biochemistry, Cardiff University School of Medicine, Heath Park, Cardiff CF14 4XN, Wales, UK. [6]Human Immunology Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA.

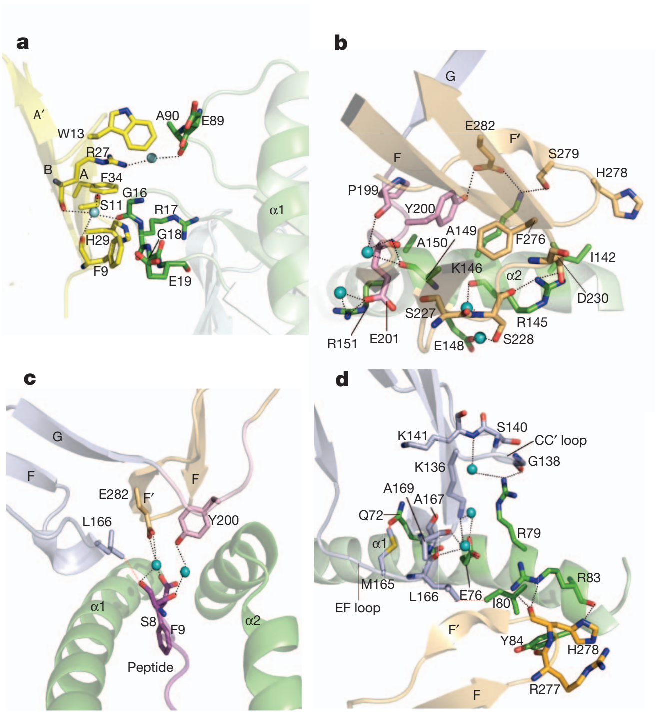


Figure 2 | Contacts between the KIR3DL1*001 receptor and pHLA-B*5701. Panels coloured as in Fig. 1. Waters are shown as cyan spheres; hydrogen bonds as black lines; van der Waals contacts as red lines. **a**, Contacts between the D0 domain and pHLA. **b**, Contacts between the receptor and the α2 helix of pHLA. **c**, Contacts to the peptide. Residues from the D1 and D2 domains form a single van der Waals and three water-mediated contacts to the P8 and P9 peptide positions. **d**, Contacts between the receptor and the α1 helix. The interface between the D1 domain and the α1 helix was suboptimal, comprising a single direct hydrogen bond from Gly 138 to Arg 79.

KIR3DL1 is 1.2 Å) (Supplementary Fig. 3a)[4,5]. Further, the positioning of the D0 domain relative to the D1 and D2 domains appears to be fixed (Supplementary Fig. 3b, c), thereby generating a pre-formed pHLA-binding receptor.

The D0 domain contributed 30% BSA upon complexation with ligand, being orientated almost perpendicular to the main axis of the antigen-binding cleft, extending towards, and just contacting, β2-microglobulin (Fig. 1a). A surface-exposed aromatic cluster (Phe 9, Trp 13, His 29, Phe 34) on one face of the D0 domain ligated to loops comprising residues 14–18 and 88–92 of HLA-B*5701 (Fig. 2a and Supplementary Table 3), both of which flexed slightly upon KIR3DL1*001 binding (Supplementary Fig. 4)[7]. These two HLA loops exhibit very limited polymorphism among the HLA-A and HLA-B allotypes and mostly have main-chain interactions with the D0 domain, thereby indicating that the D0–HLA interactions are largely independent of sequence variation and likely to be conserved across most HLA allotypes. Lengthening or shortening the HLA-B*5701 loop (residues 14–18) markedly reduced binding to KIR3DL1*001 (Fig. 3a). Alanine substitution of Ser 11, His 29 and particularly Phe 9 in KIR3DL1*001 impaired binding of HLA-B*5701 tetramers, further highlighting the importance of the D0 contacts (Fig. 3b). Interestingly, the site of the D0-mediated interaction on HLA-B*5701 has not, to the best of our knowledge, been observed in any HLA-binding immune receptor/co-receptor to date, indicating a unique molecular recognition signature, in which the D0 domain acts as an 'innate sensor' of an essentially invariant region of the HLA molecule.

The D1–D2 domains converged to form a continuous binding interface with HLA-B*5701 (Fig. 1c, d), interacting with residues from the α1- and α2-helices flanking the P8 position of the peptide. The ligand-binding site of the D1–D2 domains was relatively flat, facilitating the close positioning of HLA-B*5701, resulting in an intricate network of interactions across the interface; as such, the total BSA upon complexation at the D1–D2 interaction site was quite large (total BSA, 1,360 Å^2). The D1 and D2 domains contributed 600 and 760 Å^2 total BSA to the interface respectively, with the D1 domain docked above the α1-helix and contacting the peptide, whereas the D2 domain sat above the α2-helix, thereby providing immediate insight into the disparate roles that the D1 and D2 domains have in HLA-B*5701 engagement (Fig. 1c, d). The D2 domain predominantly interacted with a region spanning residues 142–151 of HLA-B*5701 (Supplementary Table 3), a region that shows limited polymorphism among HLA-B allotypes. At the core of the D2–HLA-B*5701 binding interface, two aromatic residues of KIR3DL1*001, Tyr 200 and Phe 276, converged onto the α2-helix, whereas polar interactions were located at the periphery (Fig. 2b). A feature of this interface was the centrally located Glu 282 of KIR3DL1*001, a charged residue that abuts Leu 166 from the D1 domain, yet is stabilized by polar interactions with Tyr 200 and Ser 279 of KIR3DL1*001, Lys 146 of HLA-B*5701 and water-mediated interactions with the peptide and Arg 83 (not shown) on the α1-helix (Fig. 2c). Alanine substitution of Glu 201, Ser 227, Asp 230 or His 278, residues that were located at the exterior of the interface, had little effect on binding (Fig. 3b). In contrast, alanine substitution of Tyr 200 or Phe 276, which formed the central aromatic cluster, or the charged residue Glu 282, abrogated tetramer binding (Fig. 3b). Further, of the five HLA-B*5701 mutations made at the D2–HLA-B*5701 interface, three residues (Ile 142, Lys 146 and Ala 149) markedly affected the affinity of the interaction (Fig. 3a). These three HLA-B*5701 residues interacted principally with Tyr 200 and Phe 276, further highlighting the importance of this internal core of KIR3DL1*001 residues in driving the D2–HLA-B*5701 interaction. Collectively, the D2–HLA-B*5701-binding site seems to have co-evolved to form a highly complementary binding interface.

KIR3DL1 recognizes HLA class I allotypes that contain the Bw4 serological epitope spanning residues 77–83 on the α1-helix[8,9].

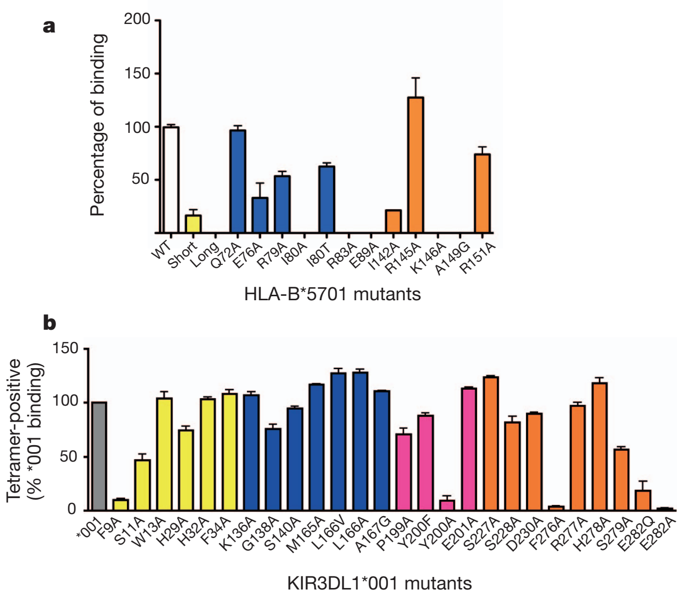


Figure 3 | Mutational analysis at the KIR3DL1*001–pHLA-B*5701 interface. **a**, Surface plasmon resonance (SPR)-based measurements of the KIR3DL1*001–HLA-B*5701 mutants interaction. Results are expressed as percentage of the wild-type interaction; mutants are colour-coded according to the KIR3DL1*001 domain they contact to correspond with Fig. 1. **b**, Capacity of HLA-B*5701 tetramers to bind 293T cells expressing wild-type or mutant KIR3DL1*001. HLA-B*5701 tetramers, but not HLA-B*0801 tetramers (data not shown), bound 293T cells transfected with KIR3DL1*001. Binding is expressed as a proportion of positive cells relative to cells transfected with wild-type KIR3DL1*001. Mutated residues are colour-coded as in Fig. 1. $N = 2$ independent experiments; error bars represent s.e.m.

While the D1 domain was positioned over the Bw4 epitope making contacts with residues 79, 80 and 83, it interacted with a broader region of the α1-helix, including Gln 72, which bound to Met 165 (Fig. 2d and Supplementary Table 3). In marked contrast to the D2-mediated contacts, the D1–HLA-B*5701 interface appeared to largely lack both charge and shape complementarity (Fig. 2d and Supplementary Fig. 5). Among the residues within the Bw4 motif, Arg 79 formed van der Waals contacts with Ser 140 and hydrogen-bonded to the main chain of Gly 138. Nevertheless the environment of Arg 79 was suboptimal, with its side chain being in close proximity to Lys 136 and Ile 139 of KIR3DL1*001. Ile 80, a residue previously associated with KIR3DL1 reactivity[10], formed a single van der Waals contact with Leu 166 and was positioned within a small hydrophobic cavity created by Glu 76, Arg 79 and Arg 83, a triad of HLA-B*5701 residues that leaned towards each other to form an array of salt-bridging interactions (Fig. 2d). Further, Arg 83 from the Bw4 motif packed against and hydrogen-bonded to the main chain of His 278 (Fig. 2d). Surprisingly, none of the five alanine mutations introduced into the D1 domain had a substantial effect on the KIR3DL1–pHLA-B*5701 interaction (Fig. 3b). However, in contrast, mutation of the corresponding HLA-B*5701 contact residues did affect recognition, particularly the Ile80Ala and Arg83Ala mutations (Fig. 3a). Interestingly, mutation of Ile 80 to Thr, a natural dimorphism within the Bw4 motif, resulted in a modest reduction in the affinity of the interaction with KIR3DL1*001 (Fig. 3a). Presumably, the Ile80Ala and Ile80Thr mutations differentially disrupt the conformation of the Glu 76-Arg 79-Arg 83 triad, thereby affecting KIR3DL1 recognition. Thus, whereas KIR3DL1*001 contacted the highly polymorphic region of the HLA class I in a non-optimal manner, and the D1 residues were shown to be non-essential for this interaction, modifications within the HLA itself affected the D1-HLA-B*5701 interaction and thus could serve to fine-tune the specificity of the interaction. Indeed, although KIR3DL1*001 specifically binds HLA molecules that possess the Bw4 motif, it does not interact with the closely related Bw6 motif, which possesses a Gly at position 83. Accordingly our data provide a basis for understanding the importance of polymorphism at residue 83 for KIR3DL1 recognition of the $Bw4^{+}$ epitope[2].

The D1 domain interacted with the LSSPVTKSF peptide; however, the sole direct interaction between the peptide and KIR3DL1*001 was a van der Waals contact between P8-Ser (where P8 is position 8 of the peptide) and Leu 166 (Fig. 2c). Thus, KIR3DL1*001 made limited contact with the peptide, analogous to the interactions observed between KIR2D and peptides bound to HLA-C[4,5], and in marked contrast to CD94-NKG2A recognition of HLA-E[11]. To probe the role of peptide in the interaction, a series of peptides that were substituted at P8 were refolded with HLA-B*5701 and assessed for their impact on recognition by KIR3DL1*001. The Phe, His and Arg P8 substitutions all facilitated an interaction with KIR3DL1*001, albeit with lower affinities, suggesting that the receptor interface has some capacity to tolerate large side chains at P8, consistent with the presence of a solvent-filled cavity adjacent to the P8 position at the KIR3DL1*001–pHLA-B*5701 interface. In contrast, the Ala, Glu and Leu P8 substitutions markedly reduced the corresponding interaction affinities (Supplementary Table 1 and Supplementary Fig. 1), suggesting that the KIR3DL1*001 receptor can 'discriminate' between peptides. The basis for the differential effects of the P8 residue could be attributable either to direct steric hindrance/lack of complementarity between the peptide and KIR3DL1*001, or to conformational alteration of the residues within the Bw4 motif itself[12]. Collectively, our observations are consistent with previous studies[13,14], which demonstrated that the sequence of the bound peptide could have a profound effect on HLA recognition by KIR.

Next, we assessed the underlying HLA specificities of the KIR2DL and KIR3DL receptor families[4,5] (Supplementary Figs 6 and 7). The D1–D2 domains of KIR3DL1*001 share clear sequence and structural homology with the HLA-C-reactive receptors, KIR2DL1, -2 and -3 (refs 4, 5, 15), and there are a number of similarities in the recognition of the α2-helix by both KIR2DL1 and -2 and KIR3DL1*001 (Supplementary Fig. 7). In contrast, the interactions between the KIR3DL1*001 and KIR2DL1 receptors and the α1-helices of their respective HLA class I ligands vary (Supplementary Fig. 6b). These differences principally arise from the loop regions that connect the C and C′ β-strands and the E and F β-strands in the D1 domain and the F and F′ β-strands that bridge the D1 and D2 domains. For example, the CC′ loop in the KIR2DL receptors adopts a notably different conformation from that observed in KIR3DL1*001 (Supplementary Fig. 6c). In KIR3DL1*001, this loop (137–140) is mostly flat and featureless, sitting adjacent to the α1-helical axis, forming limited contacts with HLA-B*5701. In the KIR2DL receptors, the corresponding loop region (42–45) is orientated towards the α1-helix, and contains two prominent residues that would prevent binding to HLA-B*5701 owing to steric hindrance with residues within the Bw4 motif. Thus, the D1-mediated contacts are critical for the HLA specificity differences between the KIR2DL1 family and KIR3DL1*001.

The KIR3D family comprises the KIR3DL1/S1, KIR3DL2 and KIR3DL3 proteins[16]. More than 200 alleles within the KIR3D family have been described, with KIR3D allomorphs generally differing from each other by a limited number of amino acids[17]. Given the high sequence identity between KIR3DL1*001 and KIR3DL2, KIR3DL3 and KIR3DS1 receptors (86, 74 and 97%, respectively), the KIR3DL1*001–HLA-B*5701 structure provided a template to examine the impact of sequence variation across the entire KIR3D family and relate this to pHLA specificity. Sequence and structural analyses suggested that a 'hotspot' resided within the D1–D2 domains, comprising loops 165–167, 199–201 and 278–282, all of which converged to form an intricate bonded network that centred on Glu 282 (Fig. 4a–e). Variation within these three loops could potentially alter the conformation of neighbouring residues within this hotspot region, thereby affecting receptor specificity.

KIR3DS1 is distinct among the KIR3D family in that it is an activating receptor. Genetic data have shown that *KIR3DL1* and *KIR3DS1* are allelic variants of the same gene and suggested that KIR3DS1 interacts with HLA-Bw4 molecules bearing an Ile at residue 80 (Bw4+I80)[18]. However, direct evidence of an interaction between KIR3DS1 and Bw4+I80 molecules is lacking[19]. Four positions that differ between KIR3DL1 and KIR3DS1 map to the KIR3DL1*001–pHLA-B*5701 interface and thus may affect the interaction (Fig. 4c), consistent with recent observations using HLA-A24 tetramers[20,21]. Whereas the Gly138Trp and Pro199Leu mutations had little impact on HLA-B*5701 binding, mutation of Leu 166—which is located within the hotspot—to Arg substantially diminished tetramer binding (Fig. 4f), thereby providing a basis for why KIR3DS1 cannot bind HLA-B*5701.

The KIR3DL2 family recognizes a limited subset of HLA-A allotypes[22,23], with seven sequence differences that map to the hotspot region (Fig. 4d). The introduction of these residues into KIR3DL1*001 showed that whereas the Leu166Pro, Ala167Val (Fig. 4f) and His278Ala mutations (Fig. 3b) did not impair recognition of HLA-B*5701, the Ser279Leu and Glu282Val mutations markedly reduced tetramer binding (Fig. 4f). Removal of the charged moiety of Glu 282 would disrupt the intricate network of interactions at the KIR3DL1*001–pHLA-B*5701 interface. Whereas the Ser279Ala mutation did not abrogate HLA-B*5701 binding (Fig. 3b), the impact of the Ser279Leu mutation was much more pronounced. This effect appears attributable to the more bulky Leu residue causing a steric clash with Arg 83, thereby suggesting a basis for the lack of reactivity of KIR3DL2 towards the Bw4 motif. Moreover, unlike HLA-B*5701 and other HLA-Bw4 allotypes, HLA-A3 and HLA-A11 possess a Gly at position 83 rather than Arg, which is a crucial determinant for KIR3DL1 recognition of the Bw4 motif[2].

The specificity of the KIR3DL3 receptor family is undefined, and a number of differences between KIR3DL1 and -3DL3 reside within the hotspot region (Fig. 4e). Binding experiments showed that the

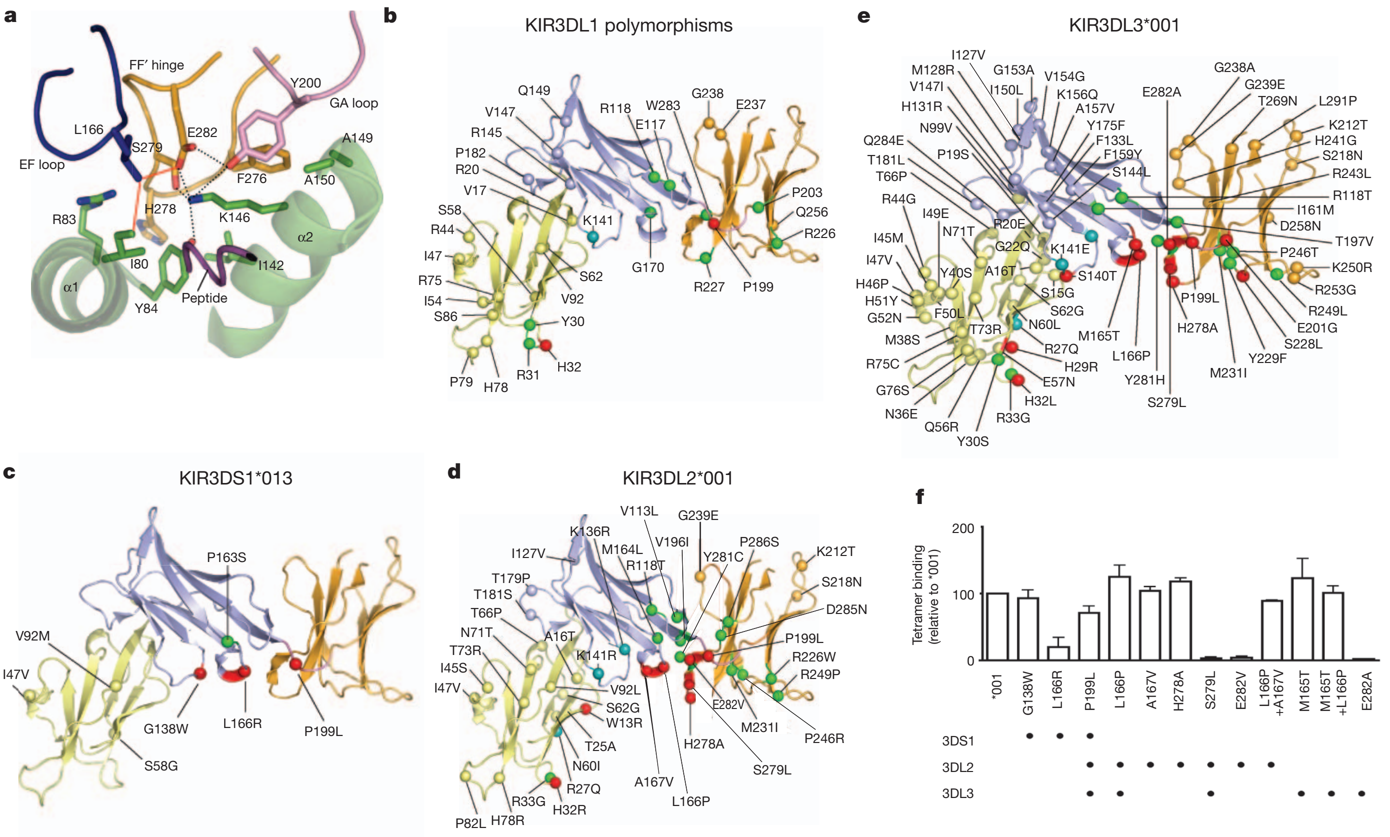


Figure 4 | Mapping of polymorphisms and sequence variations onto the structure of KIR3DL1*001. **a**, The 'hotspot' comprises three loops. **b**, Polymorphisms within the KIR3DL1 family. **c**–**e**, Differences between KIR3DS1*013 and KIR3DL1*001 (**c**), KIR3DL2*001 and KIR3DL1*001 (**d**), KIR3DL3*001 and KIR3DL1*001 (**e**). Polymorphisms are represented as spheres: red, direct contacts; cyan, water-mediated contacts; green, residues that may affect binding; other, remaining residues coloured according to domain. **f**, The capacity of HLA-B*5701 tetramers to bind to 293T cells transfected with plasmids encoding either a Flag-tagged KIR3DL1*001 or 10 site-directed mutants representing sites of 3DL1/2/3/3DS1 variation that contacted HLA-B*5701. $N = 2$ independent experiments; error bars represent s.e.m. Variations across the KIR3D family are shown underneath.

Met165Thr or Leu166Pro substitutions in KIR3DL1*001 did not affect HLA-B*5701 binding (Fig. 4f). Further, whereas the Pro199Leu substitution had a modest impact on recognition, the Glu282Ala substitution within KIR3DL1*001 totally abrogated tetramer binding (Fig. 3b), thereby indicating that residues 279 and 282 are critical determinants of the specificity differences between KIR3DL1 and other KIR3D family members.

Surprisingly, the extensive polymorphism among the inhibitory receptors within each KIR3D family was predominantly located at sites not directly implicated in pHLA binding (Fig. 4b). Collectively, these observations indicate that the majority of KIR3D polymorphisms within a family[24,25] are unlikely to directly affect the affinity of the pHLA interaction per se, but rather are likely to affect pHLA binding via altering expression levels and/or the clustering of the KIR3D receptors on the cell surface, whereas sequence differences across the KIR3D family directly affect pHLA affinity and specificity. Indeed, functional studies have shown that polymorphisms in residues such as 238 that are distant from the receptor/ligand interface can affect target cell recognition by KIR3DL1^{+} NK cells[26].

Collectively, our data provide a fundamental basis for understanding how a representative KIR3DL family member interacts with an HLA-B molecule that possesses the Bw4 motif. We show that the D0 domain, a feature of this family, interacts with a previously unrecognized determinant on the HLA molecule, which is highly conserved across HLA-A and HLA-B allotypes in particular. These observations indicate that the D0 domain acts as an innate HLA sensor at a site that that is not involved in either peptide or TCR binding[27]. The KIR3DL interaction sites seem to be largely conserved across the KIR3D family, with specificity differences mapping to a hotspot within the interaction interface. In contrast, the polymorphisms within individual *KIR3D* gene families are largely at positions that are spatially separate from the binding site, a number of which are the subject of positive selection[17]. This suggests that other evolutionary pressures, such as pathogen-mediated immune evasion strategies, may drive KIR3D diversification at sites distant from the ligand-binding site.

METHODS SUMMARY

Protein expression and purification. Inclusion body preparations of the HLA-B*5701 heavy chain and β_2-microglobulin were refolded and purified as detailed previously[7]. Residues 1–299 of KIR3DL1*001 were cloned into the pHLsec mammalian expression vector[28] with N-terminal 6×His and secretion tags. KIR3DL1*001 was expressed from transiently transfected HEK 293S cells. Purified KIR3DL1*001 was then concentrated to 15 mg ml^{-1} and deglycosylated with endoglycosidase H (New England Biolabs).

Crystallization and data collection. The KIR3DL1*001–pHLA-B*5701 complex was crystallized and its structure determined. Further details are provided in Methods.

Transfection studies. The sequence for a Flag tag (GACTACAAAGACGATGACGACAAG) was added to the 5′ end of KIR3DL1*001 by primer addition and this cDNA was then cloned into a pEF6 vector. Specific nucleotide residues were mutated using the QuikChange II Site Directed Mutagenesis Kit (Stratagene) Plasmids were transfected into HEK293T cells using the FuGene 6 transfection reagent (Roche) according to the manufacturer's instructions. After 48 h, the cells were harvested and stained with anti-Flag (clone M2, Sigma Aldrich) antibody or with tetramer for 30 min at 4 °C. The cells were then washed and analysed on a Fortessa flow cytometer (BD Biosciences).

Surface plasmon resonance. Surface plasmon resonance experiments were conducted at 25 °C on a Biacore 3000 instrument using HBS buffer (10 mM HEPES-HCl (pH 7.4), 150 mM NaCl and 0.005% surfactant P20 supplied by the manufacturer). Further details are provided in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 6 June; accepted 26 August 2011.

Published online 23 October 2011.

1. Parham, P. MHC class I molecules and KIRs in human history, health and survival. *Nature Rev. Immunol.* **5,** 201–214 (2005).
2. Sanjanwala, B., Draghi, M., Norman, P. J., Guethlein, L. A. & Parham, P. Polymorphic sites away from the Bw4 epitope that affect interaction of Bw4$^+$ HLA-B with KIR3DL1. *J. Immunol.* **181,** 6293–6300 (2008).
3. Martin, M. P. *et al.* Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nature Genet.* **39,** 733–740 (2007).
4. Boyington, J. C., Motyka, S. A., Schuck, P., Brooks, A. G. & Sun, P. D. Crystal structure of an NK cell immunoglobulin-like receptor in complex with its class I MHC ligand. *Nature* **405,** 537–543 (2000).
5. Fan, Q. R., Long, E. O. & Wiley, D. C. Crystal structure of the human natural killer cell inhibitory receptor KIR2DL1–HLA–Cw4 complex. *Nature Immunol.* **2,** 452–460 (2001).
6. Colonna, M. & Samaridis, J. Cloning of immunoglobulin-superfamily members associated with HLA-C and HLA-B recognition by human natural killer cells. *Science* **268,** 405–408 (1995).
7. Chessman, D. *et al.* Human leukocyte antigen class I-restricted activation of CD8$^+$ T cells provides the immunogenetic basis of a systemic drug hypersensitivity. *Immunity* **28,** 822–832 (2008).
8. Litwin, V., Gumperz, J., Parham, P., Phillips, J. H. & Lanier, L. L. NKB1: a natural killer cell receptor involved in the recognition of polymorphic HLA-B molecules. *J. Exp. Med.* **180,** 537–543 (1994).
9. Gumperz, J. E., Litwin, V., Phillips, J. H., Lanier, L. L. & Parham, P. The Bw4 public epitope of HLA-B molecules confers reactivity with natural killer cell clones that express NKB1, a putative HLA receptor. *J. Exp. Med.* **181,** 1133–1144 (1995).
10. Cella, M., Longo, A., Ferrara, G. B., Strominger, J. L. & Colonna, M. NK3-specific natural killer cells are selectively inhibited by Bw4-positive HLA alleles with isoleucine 80. *J. Exp. Med.* **180,** 1235–1242 (1994).
11. Petrie, E. J. *et al.* CD94-NKG2A recognition of human leukocyte antigen (HLA)-E bound to an HLA class I leader sequence. *J. Exp. Med.* **205,** 725–735 (2008).
12. Hülsmeyer, M. *et al.* Thermodynamic and structural equivalence of two HLA-B27 subtypes complexed with a self-peptide. *J. Mol. Biol.* **346,** 1367–1379 (2005).
13. Peruzzi, M., Parker, K. C., Long, E. O. & Malnati, M. S. Peptide sequence requirements for the recognition of HLA-B*2705 by specific natural killer cells. *J. Immunol.* **157,** 3350–3356 (1996).
14. Fadda, L. *et al.* Peptide antagonism as a mechanism for NK cell activation. *Proc. Natl Acad. Sci. USA* **107,** 10160–10165 (2010).
15. Maenaka, K. *et al.* Killer cell immunoglobulin receptors and T cell receptors bind peptide-major histocompatibility complex class I with distinct thermodynamic and kinetic properties. *J. Biol. Chem.* **274,** 28329–28334 (1999).
16. Trowsdale, J. *et al.* The genomic context of natural killer receptor extended gene families. *Immunol. Rev.* **181,** 20–38 (2001).
17. Norman, P. J. *et al.* Unusual selection on the KIR3DL1/S1 natural killer cell receptor in Africans. *Nature Genet.* **39,** 1092–1099 (2007).
18. Martin, M. P. *et al.* Epistatic interaction between *KIR3DS1* and *HLA-B* delays the progression to AIDS. *Nature Genet.* **31,** 429–434 (2002).
19. Carr, W. H. *et al.* Cutting edge: *KIR3DS1*, a gene implicated in resistance to progression to AIDS, encodes a DAP12-associated receptor expressed on NK cells that triggers NK cell activation. *J. Immunol.* **178,** 647–651 (2007).
20. Sharma, D. *et al.* Dimorphic motifs in D0 and D1+D2 domains of killer cell Ig-like receptor 3DL1 combine to form receptors with high, moderate, and no avidity for the complex of a peptide derived from HIV and HLA-A*2402. *J. Immunol.* **183,** 4569–4582 (2009).
21. O'Connor, G. M. *et al.* Analysis of binding of KIR3DS1*014 to HLA suggests distinct evolutionary history of KIR3DS1. *J. Immunol.* **187,** 2162–2171 (2011).
22. Dohring, C., Scheidegger, D., Samaridis, J., Cella, M. & Colonna, M. A human killer inhibitory receptor specific for HLA-A1,2. *J. Immunol.* **156,** 3098–3101 (1996).
23. Hansasuta, P. *et al.* Recognition of HLA-A3 and HLA-A11 by KIR3DL2 is peptide-specific. *Eur. J. Immunol.* **34,** 1673–1679 (2004).
24. Yawata, M. *et al.* Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection and modulation of effector function. *J. Exp. Med.* **203,** 633–645 (2006).
25. Khakoo, S. I., Geller, R., Shin, S., Jenkins, J. A. & Parham, P. The D0 domain of KIR3D acts as a major histocompatibility complex class I binding enhancer. *J. Exp. Med.* **196,** 911–921 (2002).
26. Carr, W. H., Pando, M. J. & Parham, P. KIR3DL1 polymorphisms that affect NK cell inhibition by HLA-Bw4 ligand. *J. Immunol.* **175,** 5222–5229 (2005).
27. Godfrey, D. I., Rossjohn, J. & McCluskey, J. The fidelity, occasional promiscuity, and versatility of T cell receptor recognition. *Immunity* **28,** 304–314 (2008).
28. Aricescu, A. R., Lu, W. & Jones, E. Y. A time- and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallogr. D* **62,** 1243–1250 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the staff at the MX2 beamline of the Australian synchrotron for assistance with data collection. We thank A. Radu Aricescu for the gift of the pHLsec vector. This research was supported by the National Health and Medical Research Council of Australia (NHMRC), the Australian Research Council (ARC) and the Intramural Research Programs of the National Cancer Institute and the National Institute of Allergy and Infectious Diseases, National Institutes of Health. D.W.M. and G.M.O'C. were supported by the Intramural AIDS Targeted Antiviral Program of the National Institutes of Health. J.P.V. is supported by an NHMRC Peter Doherty Research Fellowship; D.A.P. is supported by a Medical Research Council (UK) Senior Clinical Fellowship; B.A.P.L. is supported by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases, National Institutes of Health; C.S.C. is supported by an ARC QEII Fellowship; J.R. is supported by an ARC Federation Fellowship.

Author Contributions J.P.V. solved the structure, undertook analysis, performed experiments and contributed to manuscript preparation. H.H.R., T.B., R.C.D., R.B., P.M.S., M.A.O., J.M.L.W., C.M.H., J.L., S.M.M., S.G. and C.S.C. performed experiments and/or analysed data. G.M.O'C., D.A.P., B.A.P.L. and D.W.M. performed experiments and/or analysed data and contributed to the writing of the manuscript; A.G.B. and J.R. were the joint senior authors—they co-led the investigation, devised the project, analysed the data and wrote the manuscript.

Author Information The atomic coordinates and structure factors for the KIR3DL*001–pHLA-B*5701 complex were deposited in the Protein Data Bank under accession code 3VH8. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.G.B. (agbrooks@unimelb.edu.au) or J.R. (jamie.rossjohn@monash.edu).

METHODS

Protein expression and purification. HLA-B*5701 and β_2-microglobulin were expressed separately in *E. coli* from the pET-30 vector. Inclusion body preparations of the HLA-B*5701 and β_2-microglobulin were refolded and purified as detailed previously[7]. In brief, the resultant HLA class I complexes were purified by DEAE sepharose (Sigma) anion exchange chromatography using 10 mM Tris pH 8.0 and eluted with 150 mM NaCl. The protein was then further purified by gel filtration using an S200 16/60 column (GE Healthcare). The final purification step used anion exchange chromatography on a MonoQ column (GE Healthcare). The binary complex was concentrated in 10 mM Tris pH 8.0, 150 mM NaCl for use in crystallization trials and surface plasmon resonance (SPR) studies. The mutants of HLA-B*5701 were generated using the QuikChange PCR method (Stratagene) and purified as described earlier.

Residues 1–299 of KIR3DL1*001 were cloned into the pHLsec mammalian expression vector with N-terminal 6×His and secretion tags. KIR3DL1*001 was expressed from transiently transfected HEK 293S cells. Secreted KIR3DL1*001 was harvested from the culture media 3 days after transfection by first dialysing the media against 10 mM Tris pH 8.0, 300 mM NaCl before the use of nickel affinity resin. The KIR3DL1*001 was eluted from the nickel resin with 10 mM Tris pH 8.0, 300 mM NaCl, 50 mM EDTA. The protein was purified by gel filtration chromatography using an S200 16/60 column (GE Healthcare) in 10 mM Tris pH 8.0, 300 mM NaCl. Purified KIR3DL1*001 was then concentrated to 15 mg ml^{-1} and deglycosylated with endoglycosidase H (New England Biolabs.). The extent of deglycosylation was monitored by SDS–PAGE and this material was used in crystallization trials. For SPR studies a similar construct of KIR3DL1*001 was prepared in the pFastBac vector and expressed from Hi-5 insect cells (Invitrogen). The KIR3DL1*001 was purified as described earlier with the exception that the endoglycosidase H deglycosylation step was not performed.

Crystallization and data collection. The KIR3DL1*001–pHLA-B*5701 complex at 15 mg ml^{-1} was crystallized at 294 K by the hanging-drop vapour-diffusion method from a solution comprising 16% PEG 3350, 2% tacsimate pH 5 and 0.1 M tri-sodium citrate pH 5.6. The crystals typically grew to dimensions 0.3 × 0.3 × 0.2 mm in 7 days. Before data collection, the crystals were equilibrated in crystallization solution with 35% PEG 3350 added as a cryoprotectant and then flash-cooled in a stream of liquid nitrogen at 100 K. X-ray diffraction data were recorded on a Quantum-315 CCD detector at the MX2 beamline of the Australian Synchrotron. The data were integrated and scaled using DENZO and SCALEPACK from the HKL2000 program suite. Details of the data processing statistics are given in Supplementary Table 2.

The final model comprises residues 6–261, 267–292 and there are three glycosylation sites located at Asn 71, Asn 158 and Asn 252.

Structure determination and refinement. The structure was determined by molecular replacement using MOLREP. The search models used were the structures of HLA-B*5701 and KIR2DL1 (PDB codes 2RFX and 1IM9). The positions of the two complexes in the asymmetric unit were found in an incremental manner. The orientation of the first HLA molecule was found and subsequently the position of the D1 and D2 domains of the KIR receptor were placed. The second complex was fitted by application of the pseudo-translation vector 0.0, 0.5, 0.5.

Refinement of the model was carried out in REFMAC with strict twofold non-crystallographic symmetry (NCS) applied. Structure building proceeded with iterative rounds of manual building in COOT and refinement in REFMAC. The D0 domain of KIR3DL1*001 was manually built from the resultant electron density maps. The NCS restraints were removed for the final rounds of refinement. Solvent was added with COOT and the structure validated with MOLPROBITY[29]. The final structure comprises two KIR3DL1*001–pHLA-B*5701 complexes in the asymmetric unit, the association of which did not indicate higher-order oligomeric assemblies within the crystal lattice. The final refinement values are summarized in Supplementary Table 2. The crystals contained two virtually indistinguishable ternary complexes within the asymmetric unit, so structural analyses were confined to one KIR3DL1*001–pHLA-B*5701 complex.

Transfection studies. The sequence for a Flag tag (GACTACAAAGACGATGA CGACAAG) was added to the 5′ end of KIR3DL1*001 by primer addition and this cDNA was then cloned into a pEF6 vector. Specific nucleotide residues were mutated using the QuikChange II Site Directed Mutagenesis Kit (Stratagene) according to the manufacturer's instructions using PAGE-purified primers. Sequences were verified by direct sequencing. These constructs were introduced into HEK293T cells using FuGene 6 transfection reagent (Roche) according to the manufacturer's instructions. After 48 h, the cells were harvested and stained with anti-Flag (clone M2, Sigma Aldrich) antibody or with tetramer for 30 min at 4 °C. The cells were then washed and analysed on a Fortessa flow cytometer (BD Biosciences). Analysis of cell surface expression as assessed by staining with anti-Flag monoclonal antibody showed that the introduction of the mutations had no substantial effect on expression (data not shown). All transfection data are representative of two independent experiments.

SPR. SPR experiments were conducted at 25 °C on a Biacore 3000 instrument using HBS buffer (10 mM HEPES-HCl (pH 7.4), 150 mM NaCl and 0.005% surfactant P20 supplied by the manufacturer). The HLA class I-specific antibody W6/32 was immobilized on a CM5 chip via amine coupling according to manufacturer's instructions. The pHLA complexes, and mutants thereof, were captured by W6/32 creating a surface density of approximately 500–1,000 resonance units. Various concentrations of KIR3DL1*001 (2.37 to 300 μM) were injected over the captured pHLA at 5 μl min^{-1}. The final response was calculated by subtracting the response of W6/32 alone from the KIR3DL1*001–pHLA-B*5701 complex. The equilibrium data were analysed using GraphPad Prism. The shortened form of HLA-B*5701 comprised Gly-Gly-Gly in place of residues 14–19; in the long form of HLA-B*5701, Gly-Gly-Gly was inserted after Gly 16. For the SPR experiments, data are representative of two independent experiments with error bars representing s.e.m. of the duplicates.

29. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35,** W375–W383 (2007).

doi:10.1038/nature10559

Spalt mediates an evolutionarily conserved switch to fibrillar muscle fate in insects

Cornelia Schönbauer[1], Jutta Distler[2], Nina Jährling[3,4], Martin Radolf[5], Hans-Ulrich Dodt[3,4], Manfred Frasch[2] & Frank Schnorrer[1]

Flying insects oscillate their wings at high frequencies of up to 1,000 Hz[1,2] and produce large mechanical forces of 80 W per kilogram of muscle[3]. They utilize a pair of perpendicularly oriented indirect flight muscles that contain fibrillar, stretch-activated myofibres. In contrast, all other, more slowly contracting, insect body muscles have a tubular muscle morphology[4]. Here we identify the transcription factor Spalt major (Salm) as a master regulator of fibrillar flight muscle fate in *Drosophila*. *salm* is necessary and sufficient to induce fibrillar muscle fate. *salm* switches the entire transcriptional program from tubular to fibrillar fate by regulating the expression and splicing of key sarcomeric components specific to each muscle type. Spalt function is conserved in insects evolutionarily separated by 280 million years. We propose that Spalt proteins switch myofibres from tubular to fibrillar fate during development, a function potentially conserved in the vertebrate heart—a stretch-activated muscle sharing features with insect flight muscle.

To generate fast wing oscillations, both indirect flight muscle (IFM) units are attached to the thoracic exoskeleton. The contraction of one unit, the dorsal-longitudinal flight muscles (DLMs), deforms the thorax and moves the wings down; simultaneously it stretches and hence activates the second IFM unit, the dorsoventral flight muscles (DVMs), which moves the wings up again, generating an oscillatory movement of thorax and wings at high frequency[2,5]. IFMs have a unique fibrillar organization to achieve these asynchronous, stretch-activated contractions.

We performed a genome-wide RNA interference (RNAi) screen for muscle morphogenesis in *Drosophila* and identified a function for *salm* in IFM development[6]. The conserved Spalt family of transcription factors has two members in *Drosophila*, *spalt major* (*salm*) and *spalt related* (*salr*)[7]. RNAi knockdown of *salm* in muscle leads to viable but flightless animals with a reduced number of DLMs (Fig. 1a, b). Detailed analysis of the actin cytoskeleton revealed a striking change in fibre organization in *salm* knockdown IFMs: instead of the fibrillar IFM morphology with distinct, unaligned myofibrils and nuclei located between the fibrils (Fig. 1c, g and Supplementary Fig. 1a), these muscles show a tubular morphology normally found in leg muscle, with aligned myofibrils and nuclei located in the tube centre (Fig. 1d, i and Supplementary Fig. 1b). Leg muscles are normal in *salm* knockdown flies (Fig. 1e, f, h, j). We confirmed the RNAi knockdown specificity with a second independent hairpin targeting a different region of *salm* that shows an identical phenotype (data not shown) and by a small deletion that removes *salm* and its neighbouring gene *salr* (Supplementary Fig. 1c, d).

Adult muscles develop in pupae by fusion of undifferentiated adult muscle progenitors (AMPs). DLMs form by fusion of AMPs with three larval templates, inducing their splitting into the six DLMs at 14 h after pupa formation (APF) (at 27 °C)[8]. This splitting is inhibited in *salm* knockdown pupae (Supplementary Movies 1 and 2). In wild-type DLMs, myofibrils start to assemble at 30 h APF with characteristically spaced nuclei between the fibrils and distinct, unaligned fibrils visible by 45 h (Supplementary Fig. 2a–c, g–i). Leg myoblasts fuse and form tubular fibres with aligned filaments and nuclei located within the tube (Supplementary Fig. 2m–o). In *salm* knockdown IFMs, distinct fibrils never form; instead, a tubular organization similar to leg muscles develops (Supplementary Fig. 2d–f, j–l, p–r). Together, this evidence shows that *salm* is required to initiate IFM-specific muscle fate.

To investigate the mechanism of how *salm* determines IFM identity, we analysed *salm* expression. Salm is specifically expressed in adult IFMs, lost in *salm* knockdown and absent from leg muscles (Supplementary Fig. 3a–d). At 12 h APF Salm is present in the DLM templates to which the AMPs fuse. This expression increases after

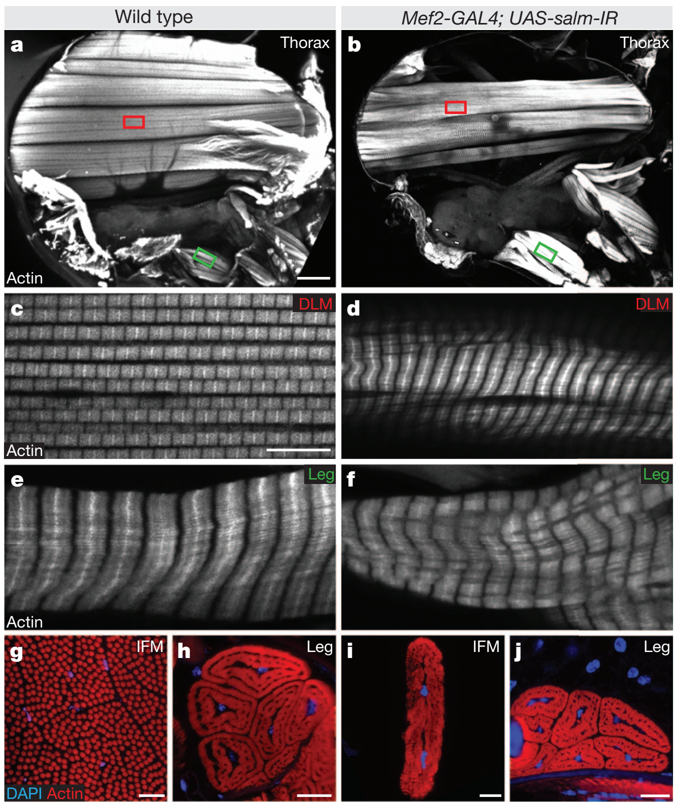


Figure 1 | *salm* specifies fibrillar flight muscle. a, b, *Drosophila* wild-type (**a**) and *Mef2-GAL4; UAS-salm-IR* (where IR is inverted repeat) (TF3029) (**b**) hemi-thorax stained with phalloidin. Boxes indicate the approximate views in **c**–**f**. **c**, **d**, Fibrillar IFMs (DLMs) in wild type (**c**) are transformed to tubular IFMs (DLMs) in *UAS-salm-IR* (**d**). **e**, **f**, Tubular leg muscles in wild type (**e**) and *UAS-salm-IR* (**f**). **g**–**j**, Cross-sections of wild-type IFMs (**g**) and leg muscles (**h**) compared to tubular IFMs (**i**) and leg muscles (**j**) in *Mef2-GAL4; UAS-salm-IR* stained with phalloidin and 4′,6-diamidino-2-phenylindole (DAPI). Scale bars 100 μm in **a**, **b**, 10 μm in **c**–**j**.

[1]Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany. [2]Friedrich-Alexander-University Erlangen-Nuremberg, Biology Department, Developmental Biology Division, Staudtstr. 5, 91058 Erlangen, Germany. [3]Vienna University of Technology, FKE, Dept. of Bioelectronics, Floragasse 7, 1040 Vienna, Austria. [4]Medical University of Vienna, Center for Brain Research, Spitalgasse 4, 1090 Vienna, Austria. [5]Institute of Molecular Pathology (IMP), Dr. Bohrgasse 7, 1030 Vienna, Austria.

template splitting at 24 h and is lost in *salm* knockdown IFMs (Fig. 2a, b, d, e and Supplementary Fig. 3e). Using a GAL4-reporter line we detect *salm* expression in the templates from 8 h APF onwards throughout IFM development (Supplementary Fig. 4 and Supplementary Movie 3). With the same line, we confirmed that *salm* is absent in developing leg muscles (Fig. 2c, f), consistent with the idea that *salm* selects fibrillar muscle fate.

If *salm* indeed specifies fibrillar muscles, overexpressing *salm* in tubular muscle should switch its sarcomere organization from tubular to fibrillar. We ectopically expressed *salm* using *Mef2-Gal4* in combination with *Tub-GAL80ts* and shifted the flies to restrictive temperature at 0 h APF, or using *1151-GAL4*, which is expressed in AMPs and developing muscles until about 40 h APF[9]. In both cases, ectopic *salm* expression induces a clear transformation of the tubular leg muscles into fibrillar IFM-like muscles (Fig. 2g–i, m, n). As a consequence, these transformed leg muscles do not function properly and flies die as pharate adults. We find a similar transformation in the abdominal muscles upon ectopic *salm* expression (Fig. 2j–l, o, p). This demonstrates that *salm* is sufficient to specify fibrillar muscle fate and to switch the developmental program from tubular to fibrillar fate. In trachea and eyes *salm* or both *salm* and *salr* are required for developmental fate decisions[10,11]. However, the selection of fibrillar flight muscle fate is largely specific to *salm*, as knockdown of *salr* by RNAi does not cause a tubular transformation, and ectopic expression of *salr* in leg or abdominal muscle does not result in a fibrillar transformation (Supplementary Fig. 5a–g). Consistently, we detect a gain of the IFM-specific protein Fln[12] and the IFM-specific isoform of Myofilin (Mf-IsoC)[13], together with a repression of the body-muscle-specific Mf-IsoB/D, Mlp84B and Mlp60 (ref. 14), in *salm*- but not in *salr*-expressing leg muscle (Supplementary Fig. 5h). Thus, we conclude that *salm* is a master regulator of *Drosophila* indirect flight muscle development.

As *salm* acts as a developmental switch, its muscle expression is restricted to IFMs. It is unclear how this precise expression is regulated. Salm is not expressed in larval AMPs (Supplementary Fig. 6a); however, the larval AMPs that build the IFMs do express the transcription factor *vestigial* (*vg*)[15] (Supplementary Fig. 6d). *vg*-null flies lack wings and halteres and have a defect in their IFMs[15]. We analysed the morphology of *vg* mutant IFMs in detail and notably found the same phenotype as in *salm* knockdown IFMs. *vg* mutant DLMs are reduced in number and show a tubular fibre phenotype (Fig. 3a, c, i). Their leg muscles are normal, which is as expected because these flies are viable and can walk (Fig. 3e). Importantly, Salm protein is lost in *vg* mutant IFMs (Fig. 3g). To investigate whether *vg* has an additional function downstream of *salm*, we expressed *salm* using *1151-GAL4* in *vg* mutants and found a complete rescue of the *vg* IFM phenotype (Fig. 3b, d, j). We did not observe a fibrillar transformation of leg muscles, possibly because Salm levels driven with *1151-GAL4* in *vg* mutant legs are too low to override the leg muscle fate (Fig. 3f, h). Interestingly, overexpression of *salr* also results in some rescue of *vg* mutant IFMs, probably mediated by regained Salm expression (Supplementary Fig. 6f, i, l, o). Together this demonstrates that *vg* is required upstream of *salm* for its IFM expression, and that *salm* does not require *vg* to implement the fibrillar flight muscle program.

Interestingly, *vg* with its cofactor *scalloped* (*sd*)[16] is not sufficient to induce fibrillar fate. Misexpression of *vg* and *sd* neither results in a fibrillar transformation nor in *salm* expression in leg muscles or wing

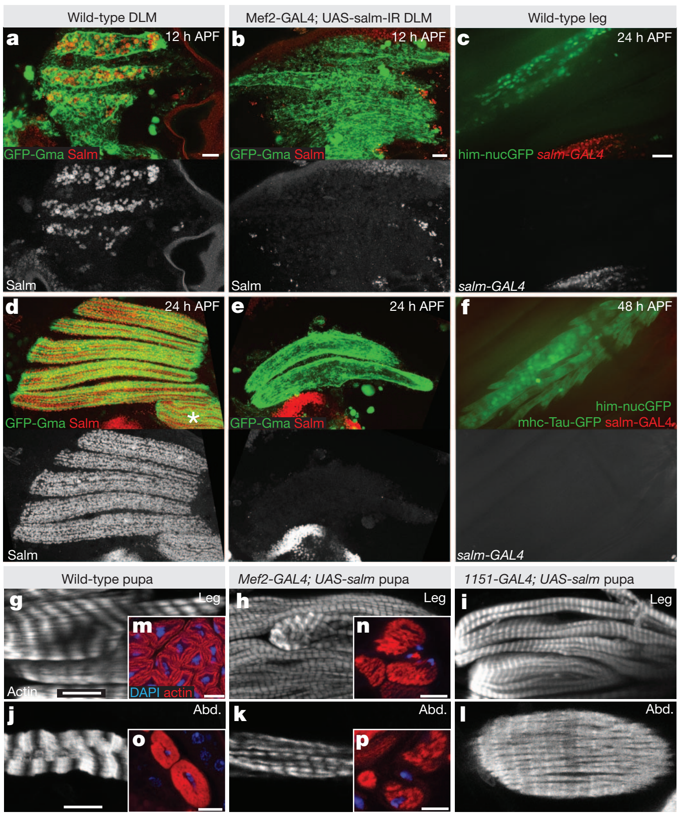


Figure 2 | Salm expression is sufficient to induce fibrillar muscle fate. **a–f**, Wild-type (**a**, **d**) or *salm* knockdown DLMs (**b**, **e**) expressing *Mef2-GAL4, UAS-GFP-gma* stained with anti-Salm at 12 h (**a**, **b**) and 24 h APF (**d**, **e**); asterisk indicates DVMs. **c**, **f**, Wild-type leg muscles labelled with *Him-nucGFP* at 24 h APF (**c**) or *Him-nucGFP* and *mhc-TauGFP* at 48 h APF (**f**). **g–l**, Phalloidin staining of late pupal leg muscles (90 h APF) (**g–i**) or abdominal muscles (**j–l**) of wild type (**g**, **j**), *Tub-GAL80ts; Mef2-GAL4; UAS-salm* shifted at 0 h APF from 18 °C to 30 °C (**h**, **k**) and *1151-GAL4; UAS-salm* (**i**, **l**). **m–p**, Cross-sections of leg and abdominal muscles of pupae with the indicated genotypes stained with phalloidin and DAPI. Scale bars, 10 µm.

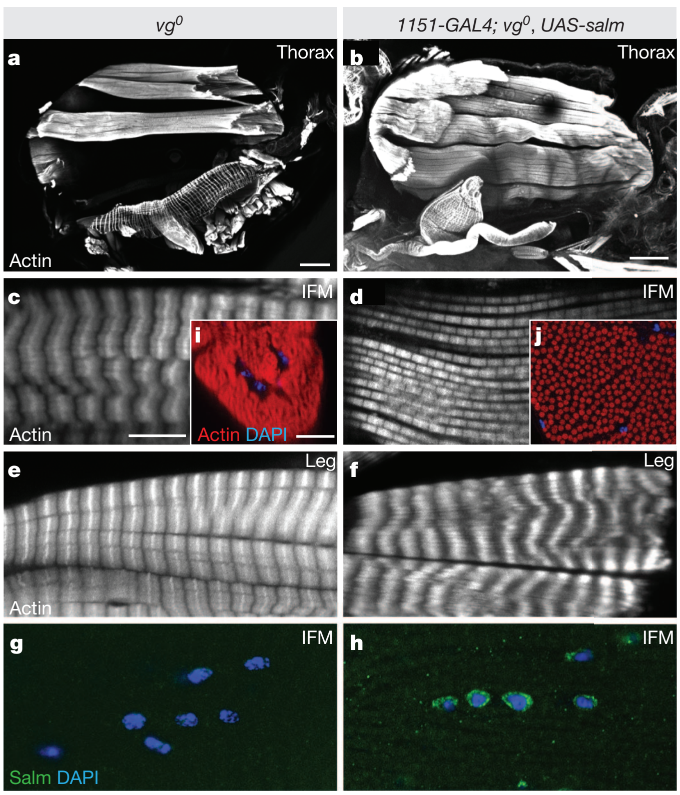


Figure 3 | *vg* functions upstream of *salm*. **a**, **b**, IFM phenotype of vg^0 mutant hemi-thorax (**a**) is rescued by expression of *UAS-salm* with *1151-GAL4* (**b**). **c–j**, IFMs in vg^0 are tubular (**c**; see **i** for cross-section), and are rescued by *1151-GAL4; UAS-salm* (**d**; see **j** for cross-section). Leg muscles are normal (**e**, **f**). Salm staining in vg^0 IFMs (**g**), *1151-GAL4;* vg^0*, UAS-salm* IFMs (**h**). Scale bars 100 µm in **a**, **b**, 10 µm in **c–h**.

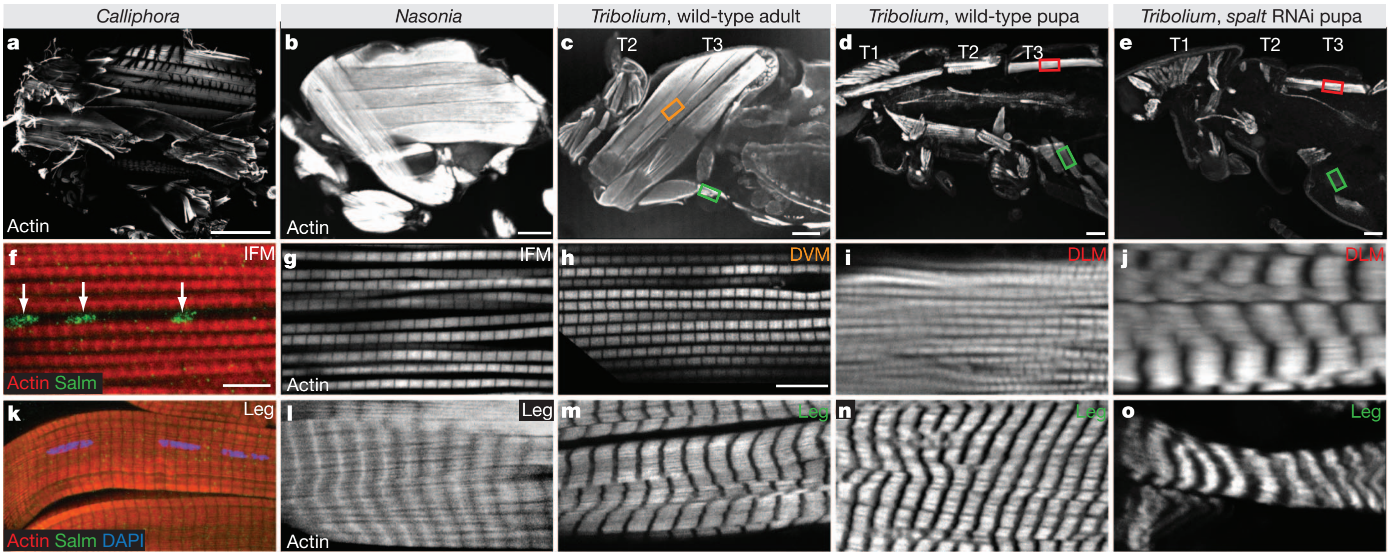


Figure 4 | Fibrillar insect flight muscle requires *spalt* function. **a**, **b**, Hemi-thorax of *Calliphora* (**a**) and *Nasonia* (**b**). **c**–**e**, Agarose sections of adult *Tribolium* (**c**), control wild-type (**d**) and *spalt* RNAi *Tribolium* pharate pupa (**e**); boxes indicate approximate areas in **h**–**j**, **m**–**o**. **f**–**o**, Indicated species or genotypes show fibrillar IFMs (**f**–**i**) and tubular leg muscles (**k**–**o**). *Calliphora* Salm is expressed in IFM nuclei (arrows, **f**) but not in leg muscles stained with DAPI (**k**). *Tribolium* DLMs are transformed from fibrillar in wild type (**i**) to tubular muscles in *spalt* RNAi (**j**). Scale bars 1 mm in **a**, 100 μm in **b**–**e**, 10 μm in **f**–**o**. The image in **a** was stitched from multiple images.

disc AMPs (Supplementary Fig. 6a–c, g, j, m, p). In contrast to *vg*, the Lbx1 homologue *ladybird early* (*lbe*) is specifically expressed in AMPs associated with the leg disc and can abrogate *vg* expression if mis-expressed in the wing disc[17] (Supplementary Fig. 6d, e). Consistently, we found that *1151-GAL4*-driven *lbe* blocks Salm expression in the IFMs, leading to tubular IFM morphology (Supplementary Fig. 6h, k, n, q). In summary, *salm*, but not *vg*, is capable of overruling the leg muscle program and determining the fibrillar muscle fate if expressed in leg myoblasts. We propose that in the absence of Salm the tubular fate program is initiated by default and does not necessarily require *lbe*, which is absent from many tubular muscles such as the abdominal muscles.

To investigate further the mechanism by which *salm* induces and executes the fibrillar program, we performed microarray analysis of dissected wild-type IFMs and *salm* knockdown IFMs using two independent hairpin constructs, and of wild-type leg muscles. Notably, we found that most known IFM-specific proteins or protein isoforms are downregulated in *salm* knockdown IFMs, including the IFM-specific stretch-sensitive TpnC4 (ref. 18), Fln[12], Mf-IsoC[13], Prm-IsoC/D[19] and Strn-Mlck-IsoE[20] (Supplementary Tables 1, 2 and Supplementary Fig. 7a). Interestingly, we also identified *vg* as downregulated, suggesting that *salm* is required to maintain *vg* expression in IFMs and initiates a feed-forward loop by activating its own activator. Consistently, *salm* knockdown leads to a gain of body-muscle-specific proteins such as MP20 (ref. 21), and body-wall-muscle-specific actins, TpnC41, Mlp84B[14], Mf-IsoB/D[13], Prm-IsoA[19] and Msp300-IsoE/G (Supplementary Table 1). The *salm*-induced switch is largely transcriptional, but also changes alternative splicing, as is the case for Mf or Strn-Mlck. We confirmed a number of these changes by western blot and antibody staining (Supplementary Fig. 7b–h). Again, *salr* expression is not changed in *salm* knockdown IFMs, arguing for a specific role of *salm* in IFM patterning. We also note that Act88F, which is enriched in IFMs as compared to leg muscles, is not changed in *salm* knockdown IFMs. However, Act88F is also expressed in a subset of tubular leg muscles, questioning its specific role in IFM development[22]. Together, these data indicate that *salm* initiates a network of gene expression by regulating transcription and alternative splicing that switches the molecular architecture of the muscle from tubular to fibrillar morphology.

Many winged insects use IFMs to move their wings at various frequencies[1,2]. We wished to determine the IFM morphology in different insect orders across an evolutionary distance of 280 million years (Supplementary Fig. 8)[23]. We chose *Calliphora* as a second dipteran species, the wasp *Nasonia* as a hymenopteran and the beetle *Tribolium* as a coleopteran representative. All these species have a fibrillar organization of their IFMs and a tubular organization of their leg muscles (Fig. 4a–c, f–h, k–o). We found that Salm expression in *Calliphora* is IFM specific (Fig. 4f, k), indicating that the functional distinction of muscle types correlates with *salm* expression in dipteran species. To investigate functionally a potential role of *spalt*, we used systemic RNAi in *Tribolium*[24]. Injection of *spalt* dsRNA into *Tribolium* larvae leads to pupae that are unable to complete metamorphosis and die as pharate adults. Histological analysis of the DLMs reveals a marked transformation to the tubular muscle morphology after *spalt* knockdown (Fig. 4e, j), as opposed to the fibrillar morphology in control injected animals (Fig. 4d, i). Hence, *spalt* is required in *Tribolium*, as it is in *Drosophila*, to specify fibrillar flight muscles, suggesting that *spalt* function as a regulator of fibrillar flight muscles is conserved in all insects harbouring stretch-activated indirect flight muscles.

Mice and humans possess four *spalt-like* (*SALL*) genes, none of which are expressed in differentiated striated body muscles[25,26]. This is not surprising, as all vertebrate body muscles harbour aligned sarcomeres that resemble the tubular insect muscles. Interestingly, *SALL1* and *SALL3* are both expressed in mouse and human hearts[25,26], which contain distinct unaligned myofibrils in cardiomyocytes[27] and utilize the stretch-modulated Frank–Starling contraction mechanism[28]. Mutations in human *SALL1* cause the heart abnormalities observed in Townes–Brocks syndrome[29], leading us to speculate that *spalt* function determines fibrillar stretch-activated muscle all the way up to vertebrates.

METHODS SUMMARY

All RNAi crosses were performed at 27 °C. Adult or pupal flight or leg muscles were bisected and stained with phalloidin or antibodies. For early pupal stages muscles were dissected or pupae were embedded in agarose and sectioned. For time-lapse movies pupae were mounted in Voltalef oil and imaged using a spinning disc confocal.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 May; accepted 13 September 2011.

1. Dudley, R. in *The Biomechanics of Insect Flight* (ed. Dudley, R.) 75–158 (Princeton Univ. Press, 2000).
2. Dickinson, M. Insect flight. *Curr. Biol.* **16,** R309–R314 (2006).

3. Lehmann, F. O. & Dickinson, M. H. The changes in power requirements and muscle efficiency during elevated force production in the fruit fly *Drosophila melanogaster*. *J. Exp. Biol.* **200,** 1133–1143 (1997).
4. Bernstein, S. I., O'Donnell, P. T. & Cripps, R. M. Molecular genetic analysis of muscle development, structure, and function in *Drosophila*. *Int. Rev. Cytol.* **143,** 63–152 (1993).
5. Dudley, R. in *The Biomechanics of Insect Flight* (ed. Dudley, R.) 36–74 (Princeton Univ. Press, 2000).
6. Schnorrer, F. *et al.* Systematic genetic analysis of muscle morphogenesis and function in *Drosophila*. *Nature* **464,** 287–291 (2010).
7. de Celis, J. F. & Barrio, R. Regulation and function of Spalt proteins during animal development. *Int. J. Dev. Biol.* **53,** 1385–1398 (2009).
8. Dutta, D., Anant, S., Ruiz-Gomez, M., Bate, M. & VijayRaghavan, K. Founder myoblasts and fibre number during adult myogenesis in *Drosophila*. *Development* **131,** 3761–3772 (2004).
9. Anant, S., Roy, S. & VijayRaghavan, K. Twist and Notch negatively regulate adult muscle differentiation in *Drosophila*. *Development* **125,** 1361–1369 (1998).
10. Franch-Marro, X. & Casanova, J. *spalt*-induced specification of distinct dorsal and ventral domains is required for *Drosophila* tracheal patterning. *Dev. Biol.* **250,** 374–382 (2002).
11. Mollereau, B. *et al.* Two-step process for photoreceptor formation in *Drosophila*. *Nature* **412,** 911–913 (2001).
12. Reedy, M. C., Bullard, B. & Vigoreaux, J. O. Flightin is essential for thick filament assembly and sarcomere stability in *Drosophila* flight muscles. *J. Cell Biol.* **151,** 1483–1500 (2000).
13. Qiu, F. *et al.* Myofilin, a protein in the thick filaments of insect muscle. *J. Cell Sci.* **118,** 1527–1536 (2005).
14. Stronach, B. E., Siegrist, S. E. & Beckerle, M. C. Two muscle-specific LIM proteins in *Drosophila*. *J. Cell Biol.* **134,** 1179–1195 (1996).
15. Bernard, F. *et al.* Control of *apterous* by *vestigial* drives indirect flight muscle development in *Drosophila*. *Dev. Biol.* **260,** 391–403 (2003).
16. Halder, G. *et al.* The Vestigial and Scalloped proteins act together to directly regulate wing-specific gene expression in *Drosophila*. *Genes Dev.* **12,** 3900–3909 (1998).
17. Maqbool, T. *et al.* Shaping leg muscles in *Drosophila*: role of *ladybird*, a conserved regulator of appendicular myogenesis. *PLoS ONE* **1,** e122 (2006).
18. Agianian, B. *et al.* A troponin switch that regulates muscle contraction by stretch instead of calcium. *EMBO J.* **23,** 772–779 (2004).
19. Arredondo, J. J. *et al.* Control of *Drosophila* paramyosin/miniparamyosin gene expression. Differential regulatory mechanisms for muscle-specific transcription. *J. Biol. Chem.* **276,** 8278–8287 (2001).
20. Patel, S. R. & Saide, J. D. Stretchin-klp, a novel *Drosophila* indirect flight muscle protein, has both myosin dependent and independent isoforms. *J. Muscle Res. Cell Motil.* **26,** 213–224 (2005).
21. Ayme-Southgate, A., Lasko, P., French, C. & Pardue, M. L. Characterization of the gene for mp20: a *Drosophila* muscle protein that is not found in asynchronous oscillatory flight muscle. *J. Cell Biol.* **108,** 521–531 (1989).
22. Nongthomba, U., Pasalodos-Sanchez, S., Clark, S., Clayton, J. D. & Sparrow, J. C. Expression and function of the *Drosophila* ACT88F actin isoform is not restricted to the indirect flight muscles. *J. Muscle Res. Cell Motil.* **22,** 111–119 (2001).
23. Savard, J. *et al.* Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of holometabolous insects. *Genome Res.* **16,** 1334–1338 (2006).
24. Tomoyasu, Y. & Denell, R. E. Larval RNAi in *Tribolium* (Coleoptera) for analyzing adult development. *Dev. Genes Evol.* **214,** 575–578 (2004).
25. Parrish, M. *et al.* Loss of the *Sall3* gene leads to palate deficiency, abnormalities in cranial nerves, and perinatal lethality. *Mol. Cell. Biol.* **24,** 7102–7112 (2004).
26. Nishinakamura, R. *et al.* Murine homolog of *SALL1* is essential for ureteric bud invasion in kidney development. *Development* **128,** 3105–3115 (2001).
27. Manisastry, S. M., Zaal, K. J. & Horowits, R. Myofibril assembly visualized by imaging N-RAP, α-actinin, and actin in living cardiomyocytes. *Exp. Cell Res.* **315,** 2126–2139 (2009).
28. Shiels, H. A. & White, E. The Frank–Starling mechanism in vertebrate cardiac myocytes. *J. Exp. Biol.* **211,** 2005–2013 (2008).
29. Surka, W. S., Kohlhase, J., Neunert, C. E., Schneider, D. S. & Proud, V. K. Unique family with Townes–Brocks syndrome, *SALL1* mutation, and cardiac defects. *Am. J. Med. Genet.* **102,** 250–257 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Affolter, D. Bäumer, M. Beckerle, B. Bullard, E. Chen, K. Clark, C. Desplan, K. Jagla, A. Lalouette, J. Posakony, D. Reiff, J. Saide, S. Sprecher, R. Schuh, G. Tanentzapf, J. Vigoreaux, K. VijayRaghavan, the Bloomington and the VDRC stock centres for fly stocks, antibodies and insect species. We are grateful to B. Dickson, I. Hein, M. Klingler and M. Sixt for discussions, and to R. Fässler for support and discussions. We thank A. Kaya-Copur, H. Knaut, M. Spletter and N. Vogt for critical comments on the manuscript. This work was supported by the Max-Planck-Society, a Career Development Award by the Human Frontier Science Programme to F.S., a Doc-fForte predoctoral fellowship from the Austrian Academy of Sciences to C.S., and DFG grants to M.F.

Author Contributions C.S. performed most of the experiments, analysed the data and created most of the figures. F.S. acquired the time-lapse movies and performed western blots. J.D. and M.F conducted the Tribolium RNAi experiments, M.R. performed the microarray analysis, and N.J. and H.-U.D. were involved in the initial characterisation of the *salm* mutant phenotype. F.S. conceived and supervised the project and wrote the manuscript with input from C.S. and M.F.

Author Information All raw data were submitted to the Gene Expression Omnibus under accession number GSE27502. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to F.S. (schnorrer@biochem.mpg.de).

METHODS

Fly strains and genetics. All fly work, unless otherwise stated, was performed at 27 °C to enhance GAL4 activity. Two independent *UAS-salm-IR* lines (TF3029 and TF101052) were obtained from the VDRC stock centre. Sequences for all VDRC RNAi hairpins are deposited at http://stockcenter.vdrc.at; TRiP sequences can be found at http://www.flyrnai.org. *salm* hairpins were crossed to *Mef2-GAL4* (ref. 30). For knockdown of *salr*, we used TF28386 from the VDRC stock centre and the JF03226 TRiP line driven with *1151-GAL4* (ref. 9) or *Mef2-GAL4*, respectively. For ectopic expression of *salm, UAS-salm*[31] was crossed to *Tub-GAL80ts; Mef2-GAL4* or *1151-GAL4* at 18 °C and shifted to 30 °C at 0 h APF to prevent early lethality. Similarly, crosses of *UAS-salm,* vg^0 (ref. 15) with *1151-GAL4;* vg^0*/CyO* were kept at 18 °C until 0 h APF and then shifted to 30 °C. Misexpression of *UAS-lbe* and *UAS-vg* was performed with *1151-GAL4* at 25 °C. For *salm* expression during pupal development, we used flies expressing *salm-GAL4/CyO* and *him-nuclear-GFP,* which marks undifferentiated myoblasts[32], and with *mhc-TauGFP*[33] labelling all differentiated muscles. For the *salm* mutant mitotic clones, the FRT cell-lethal method was used[34]. *hs-Flp; Df(2L)32FP-5, FRT40A/cl2L3, FRT40A* larvae were grown at 25 °C and heat-shocked twice for 60 min at 37 °C on two consecutive days. The IFM phenotype of flightless animals was analysed as described below. To construct *UAS-salr* the 8.0-kb *salr* genomic region was amplified with gene-specific primers (tcgtagtcaagttcgttccagg and ttgtgtcagtgctagta gaagc) from genomic DNA and cloned into pUAST. Transgenic lines were generated using standard procedures.

Analysis of IFMs and leg muscles. Hemi-thoraces for imaging *Drosophila* adult and pharate pupal IFM and leg muscles were prepared and stained as described[6]. Actin was visualized with rhodamine phalloidin (Molecular Probes). Rabbit anti-Fln[12] and rabbit anti-Salm[35] were used at 1:50, and rabbit anti-Mlp84B was diluted 1:500[14]. Nuclei were visualized with DAPI or mouse anti-Lamin (Hybridoma Bank, clone ADL67.10) was used at 1:10. *Calliphora* IFM and leg muscle morphology was analysed by bisection of thoraces and staining with Salm antibody, phalloidin and DAPI. To examine the *salm* knockdown phenotype during late pupal development (30–60 h APF), wild-type *Mef2-GAL4, UAS-GFP-Gma* and *Mef2-GAL4, UAS-salm-IR, UAS-GFP-Gma* pupae of desired stages were fixed in 4% paraformaldehyde in PBST (with 0.5% Triton-X 100) overnight at 4 °C, washed twice for 10 min in PBST and then embedded in 7% agarose. Agarose blocks were cut in 90-μm sections with a vibratome. Sections were incubated with DAPI for 10 min, washed twice for 10 min in PBST and mounted in Vectashield. Similarly, *Tribolium,* and *Nasonia* IFMs, leg muscles, and all cross-sections were analysed by staining agarose sections with rhodamine phalloidin and DAPI. To analyse IFM morphology and *salm* expression in wild-type and *salm* knockdown flies during early pupal development, 12 h and 24 h APF pupae were dissected as described[36].

Immunolabelling of larval imaginal discs. Dissection and staining of 3rd larval instar wing and leg discs was performed as described[37]. Wing and leg discs associated AMPs were labelled with *1151-GAL4, UAS-GFP-gma* and rabbit anti-Salm at 1:50 or anti-Vg at 1:200.

Time-lapse movies. Staged 8–10 h pupae were carefully cleaned with a wet brush and transferred into a custom-made slide with a slit fitting an entire pupa, dorsal side facing up. The pupa was slightly turned (10–20°) resulting in DLM templates facing up. A coverslip with a thin layer of 3S Voltalef oil facing the pupa was placed on top. Z-stack images were acquired every 5 min using a spinning disc confocal with a ×20 or ×40 objective (Zeiss, Visitron).

RNAi in *Tribolium*. A 3,320-bp Tc'spalt (BeetleBase TC013501; GenBank CM000280.2) fragment was amplified from cDNA with gene-specific primers (Tc'sal P2 5′-CACCCTCCAGCACCAACAAG-3′, Tc'sal P7 5′-CCCCGTTG CTCCACATATGC-3′) and cloned into pBluescript 2. PCR templates from this clone were generated with a T7 and a fused T7–T3 primer and used for *in vitro* dsRNA synthesis with the MEGAscript T7 High Yield Transcription Kit (Ambion). dsRNA injections of 4th and 5th instar larvae (*Tribolium* wild-type strain San Bernardino) were performed as described[24]. Larvae were anaesthetized on ice for 15 min and abdominally injected with *spalt* dsRNA (1 μg μl^{-1}) until the larvae had stretched visibly. After injection the beetles were kept on flour (5% yeast, 0.5% fumagillin) at 32 °C.

Microarray analysis. Wild-type IFMs, *salm* knockdown IFMs and leg muscles were dissected in PBS and homogenized in TriPure (Roche). RNA was extracted, labelled and hybridized to Agilent microarrays according to the manufacturer (Agilent). All experiments were performed in biological duplicates with one additional technical replicate. Log_2 fold change ratios of genes expressed above threshold 8.5 were averaged. All raw data were submitted to the Gene Expression Omnibus (GSE27502).

Western blot. Protein extracts from adult thoraces (without wings and legs), entire legs or dissected IFMs were blotted using standard procedures. Rabbit anti-Fln[12], rabbit anti-Mf[13], and rabbit anti-Mlp60 (ref. 14) were used at 1:10,000, and rabbit anti-Mlp84B at 1:20,000 (ref. 14).

30. Ranganayakulu, G., Schulz, R. A. & Olson, E. N. Wingless signaling induces *nautilus* expression in the ventral mesoderm of the *Drosophila* embryo. *Dev. Biol.* **176,** 143–148 (1996).
31. Grieder, N. C., Morata, G., Affolter, M. & Gehring, W. J. *Spalt major* controls the development of the notum and of wing hinge primordia of the *Drosophila melanogaster* wing imaginal disc. *Dev. Biol.* **329,** 315–326 (2009).
32. Rebeiz, M., Reeves, N. L. & Posakony, J. W. SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl Acad. Sci. USA* **99,** 9888–9893 (2002).
33. Chen, E. H. & Olson, E. N. Antisocial, an intracellular adaptor protein, is required for myoblast fusion in *Drosophila. Dev. Cell* **1,** 705–715 (2001).
34. Newsome, T. P., Asling, B. & Dickson, B. J. Analysis of *Drosophila* photoreceptor axon guidance in eye-specific mosaics. *Development* **127,** 851–860 (2000).
35. Kuhnlein, R. P. *et al. spalt* encodes an evolutionarily conserved zinc finger protein of novel structure which provides homeotic gene function in the head and tail region of the *Drosophila* embryo. *EMBO J.* **13,** 168–179 (1994).
36. Fernandes, J. J., Celniker, S. E. & VijayRaghavan, K. Development of the indirect flight muscle attachment sites in *Drosophila*: role of the PS integrins and the stripe gene. *Dev. Biol.* **176,** 166–184 (1996).
37. Klein, T. in Drosophila*: Methods and Protocols* (ed. Dahmann, C.) 253–264 (Humana Press, 2008).

doi:10.1038/nature10607

A heteromeric Texas coral snake toxin targets acid-sensing ion channels to produce pain

Christopher J. Bohlen[1]*, Alexander T. Chesler[1]*, Reza Sharif-Naeini[2], Katalin F. Medzihradszky[3], Sharleen Zhou[4], David King[4], Elda E. Sánchez[5], Alma L. Burlingame[3], Allan I. Basbaum[2] & David Julius[1]

Natural products that elicit discomfort or pain represent invaluable tools for probing molecular mechanisms underlying pain sensation[1]. Plant-derived irritants have predominated in this regard, but animal venoms have also evolved to avert predators by targeting neurons and receptors whose activation produces noxious sensations[2–6]. As such, venoms provide a rich and varied source of small molecule and protein pharmacophores[7,8] that can be exploited to characterize and manipulate key components of the pain-signalling pathway. With this in mind, here we perform an unbiased *in vitro* screen to identify snake venoms capable of activating somatosensory neurons. Venom from the Texas coral snake (*Micrurus tener tener*), whose bite produces intense and unremitting pain[9], excites a large cohort of sensory neurons. The purified active species (MitTx) consists of a heteromeric complex between Kunitz- and phospholipase-A2-like proteins that together function as a potent, persistent and selective agonist for acid-sensing ion channels (ASICs), showing equal or greater efficacy compared with acidic pH. MitTx is highly selective for the ASIC1 subtype at neutral pH; under more acidic conditions (pH < 6.5), MitTx massively potentiates (>100-fold) proton-evoked activation of ASIC2a channels. These observations raise the possibility that ASIC channels function as coincidence detectors for extracellular protons and other, as yet unidentified, endogenous factors. Purified MitTx elicits robust pain-related behaviour in mice by activation of ASIC1 channels on capsaicin-sensitive nerve fibres. These findings reveal a mechanism whereby snake venoms produce pain, and highlight an unexpected contribution of ASIC1 channels to nociception.

To identify novel toxins that activate nociceptors, we screened venoms from a variety of snake species for their ability to depolarize specific subpopulations of somatosensory neurons using calcium imaging as a functional readout. Among these, venom from the Texas coral snake, which elicits intense acute pain associated with local oedema and inflammation[9], produced clear and robust activation of most, but not all, neurons cultured from trigeminal ganglia of newborn (P0–P2) rats (Fig. 1a, b). In contrast, coral snake venom had no effect on sympathetic neurons cultured from superior cervical ganglia (not shown).

We next fractionated the crude venom using reversed-phase chromatography (Supplementary Fig. 1) to identify the active component(s). No single fraction excited sensory neurons, but when re-pooled the individual fractions fully reconstituted activity observed with the crude venom, suggesting a requirement for multiple components. Pair-wise analysis of these fractions identified two components (MitTx-α and MitTx-β), which could be purified to near-homogeneity with subsequent chromatographic steps, as assessed by gel filtration chromatography, SDS–polyacrylamide gel electrophoresis and mass spectrometry (not shown). These two toxins together proved necessary and sufficient to recapitulate activity of the crude venom. Mass spectrometry and amino (N)-terminal Edman sequencing showed that both MitTx-α and MitTx-β are proteinaceous in nature, and partial amino-acid sequences derived from these analyses were used to clone full-length complementary DNAs (cDNAs) from the coral snake venom gland. The deduced amino-acid sequences revealed patterns of cysteine residues that classify MitTx-α and MitTx-β as Kunitz type and phospholipase A2 (PLA2)-like proteins, respectively (Fig. 1c and Supplementary Fig. 1). Indeed, these families of cysteine-rich disulphide-bonded proteins are prevalent components of various snake venoms[8], and in some cases have been shown to form biochemical complexes[10].

To determine whether MitTx-α and MitTx-β form a heteromeric complex, we used isothermal titration calorimetry to detect any such molecular interaction. We observed a substantial exothermic reaction (change in enthalpy $\Delta H = -18.9 \pm 1.3\,\mathrm{kcal\,mol^{-1}}$) upon mixing, resulting from a high-affinity binding event (dissociation constant $K_d = 12.2 \pm 3.1$ nM) with 1:1 stoichiometry ($n = 1.02 \pm 0.05$) (Fig. 1d). These biochemical results are consistent with our physiological analysis showing that neither MitTx-α nor MitTx-β activated sensory neurons alone, whereas a robust and immediate rise in intracellular calcium was produced when both components were added simultaneously or sequentially (in either order) (Fig. 1e). Moreover, even a brief washout period between sequential applications prevented activation (Fig. 1f), suggesting that only the MitTx-α/β complex forms a persistent and productive interaction with its physiological target.

Having defined the molecular nature of the toxin components, we next sought to elucidate its mechanism of action on sensory neurons. As the MitTx-β component lacks critical catalytic residues normally found in the active site of related PLA2 enzymes[11], phospholipase activity seemed unlikely to contribute to neuronal excitation. In fact, we failed to detect PLA2 activity associated with any component, together or alone (Supplementary Fig. 1), suggesting that the toxin is not producing neuronal depolarization simply by degrading the plasma membrane (consistent with the observed action on a subset of somatosensory neurons). Nevertheless, it is possible that MitTx-β maintains some lipid-binding character from its PLA2 lineage, which could aid in effecting neuronal depolarization.

To gain clues about the cellular mechanism of toxin action, we performed a detailed biophysical analysis of toxin-evoked responses. Whole-cell patch-clamp recordings showed that purified MitTx-α/β complex (hereafter referred to as MitTx) produced robust currents in a subset of trigeminal neurons. These currents were characterized by a linear current–voltage relationship and high permeability to Na^+ versus Cs^+ ions ($P_{Na^+}:P_{Cs^+} = 10.1 \pm 1.0$, Fig. 2a). In searching for potential molecular targets expressed by sensory neurons, we next investigated members of the ASIC family, which exhibit properties consistent with these parameters[12–14]. We expressed various ASIC subtypes in *Xenopus* oocytes and measured toxin-evoked electrophysiological responses. Figure 2 shows that application of MitTx produced large and sustained

[1]Department of Physiology, University of California, San Francisco, California 94158-2517, USA. [2]Department of Anatomy, University of California, San Francisco, California 94158-2517, USA. [3]Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158-2517, USA. [4]Howard Hughes Medical Institute Mass Spectrometry Laboratory, University of California, Berkeley, California 94720-3202, USA. [5]National Natural Toxins Research Center and Department of Chemistry, Texas A&M University-Kingsville, Texas 78363, USA.
*These authors contributed equally to this work.

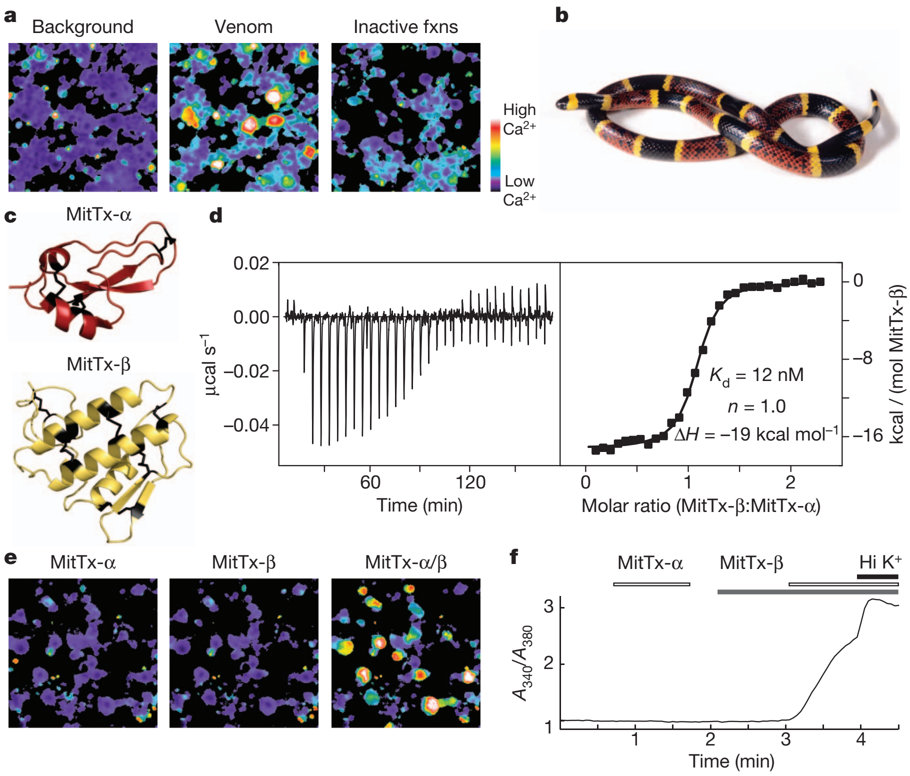


Figure 1 | Heteromeric toxin from Texas coral snake activates somatosensory neurons. a, *M. t. tener* venom (0.1 mg ml^{-1}) activates acutely dissociated trigeminal ganglion neurons as assessed by ratiometric calcium imaging. Pooled venom fractions lacking neuron-specific activity (inactive fxns) produced only weak signals in non-neuronal cells (colour bar indicates relative calcium levels). **b**, The Texas coral snake. **c**, Homology-based predicted structural models of MitTx subunits, generated using Prime (Schrodinger)[29]. **d**, Isothermal titration calorimetry reveals formation of high-affinity MitTx-α/β complex with 1:1 stoichiometry. **e**, MitTx-α and MitTx-β have no effect individually, but recapitulate activity of whole venom when applied together to trigeminal ganglia neurons. **f**, Average calcium response from more than 100 randomly selected trigeminal ganglia neurons that also responded to high extracellular potassium (Hi K$^+$, 100 mM KCl). Activation by MitTx-α/β (300 nM each) only occurs when both toxins are present.

membrane currents in oocytes expressing the ASIC1b subtype (Fig. 2b, c). Consistent with our results using cultured neurons, neither MitTx-α nor MitTx-β was active on its own, but robust responses were evoked when both components were applied to oocytes, whether pre-mixed or mixed *in situ*. Furthermore, the ENaC/ASIC blocker, amiloride, abolished these responses. When applied at maximal concentration, MitTx elicited responses exceeding those produced by saturating doses of extracellular protons. Whereas protons elicit very transient responses, those evoked by toxin were dramatically prolonged, reflecting both lack of desensitization and slow reversibility after washout (Fig. 2c).

Among ASIC subtypes, the most robust toxin-evoked responses were observed with ASIC1a or 1b based on potency (half-maximum effective concentration $EC_{50} = 9.4 \pm 1.3$ and 23 ± 3.6 nM, respectively), efficacy (relative to protons) and persistence of action (Fig. 2d and Supplementary Fig. 2). MitTx must interact with an extracellular region of the channel because responses were observed in the outside-out (but not inside-out) configuration when toxin was applied to patches excised from ASIC1a-expressing CHO cells (not shown). Moreover, at least for ASIC1a, toxin potency (9 nM) is probably limited by the affinity of α/β complex formation (12 nM, Fig. 1d) and not by the toxin–channel interaction itself.

The ASIC3 subtype was also MitTx sensitive, but required ~100-fold higher toxin concentration to achieve appreciable activation ($EC_{50} = 830 \pm 250$ nM). This lower potency was accompanied by relatively slow activation and fast washout kinetics compared with those observed with ASIC1a or 1b (Supplementary Fig. 2). By comparison, ASIC2a showed very weak activation by toxin, never achieving more than 10% efficacy compared with proton-evoked responses (Fig. 2d, e). Despite this anaemic response, the toxin produced a remarkable potentiation of acid-evoked currents, greatly enhancing both potency and efficacy of protons (Fig. 2e, f). Indeed, as the extracellular pH drops below neutrality, the toxin itself becomes an excellent ASIC2a agonist, essentially enhancing the potency of protons by three orders of magnitude. Two remaining ASIC family members, 2b and 4, do not produce proton-activated channels on their own, and we did not observe MitTx-evoked responses in oocytes expressing these subtypes. As a further testament to toxin specificity, MitTx produced neither activation nor persistent inhibition when applied to a diverse range of cloned ion channels, including voltage-gated, ENaC, TRP, P2X or 5-HT_3 channels (Supplementary Fig. 3). Additionally, MitTx activated the same percentage of trigeminal neurons cultured from wild-type or TRPV1/TRPM8/TRPA1 triple knockout mice (Supplementary Fig. 4).

M. tener tener is not the only coral snake species to express ASIC-activating toxins; venom from the Brazilian coral snake (*Micrurus frontalis*) activated a similar cohort of cultured rat trigeminal sensory neurons, or oocytes expressing cloned ASIC1a channels (Supplementary Fig. 5). Interestingly, ASIC1a is also targeted by a peptide toxin from tarantula (PcTx1), but in this case, the toxin serves as a functional antagonist of proton-evoked responses by locking the channel in a desensitized state[15,16]. ASIC1a channels blocked by PcTx1 could not be activated by MitTx, and MitTx-activated channels could not be blocked by PcTx1 (Fig. 2g, h), suggesting physical or functional occlusion of toxin action.

We next used patch-clamp recording methods to determine whether MitTx-evoked neuronal responses exhibit properties consistent with an ASIC-mediated mechanism. Sensory neurons show both transient and sustained proton-evoked responses, with the former being mediated primarily by ASIC channels and the latter by capsaicin-sensitive TRPV1 channels[17,18]. We found that all MitTx-sensitive neurons exhibited transient responses to extracellular protons (pH 4), toxin responses were blocked by amiloride and eliminated in Na$^+$-free

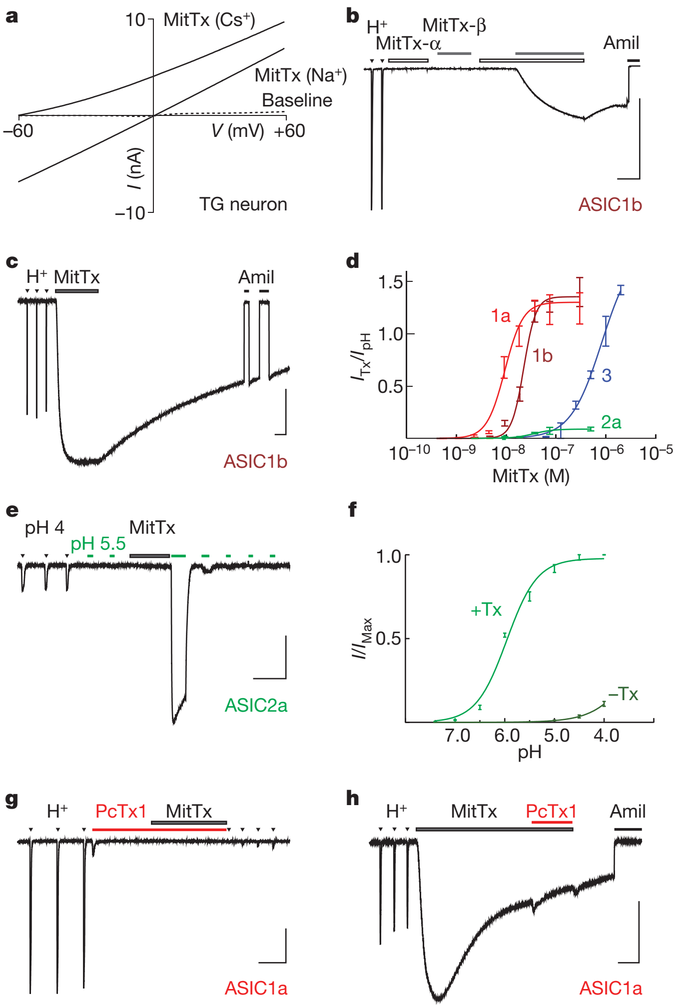


Figure 2 | MitTx activates ASICs. **a**, Current–voltage relationships of MitTx (300 nM)-evoked conductances from trigeminal ganglia neurons (whole-cell configuration) demonstrate higher permeability for Na^+ over Cs^+. The intracellular solution contained 150 mM Na^+, and a leftward shift in the reversal potential was observed when the major extracellular cation was changed from 150 mM Na^+ to 150 mM Cs^+. **b**, Voltage-clamp recordings show that ASIC1b-expressing oocytes respond to both extracellular acidification (H^+, pH 4) and MitTx, but are insensitive to MitTx-α (30 nM) or MitTx-β (300 nM) individually. Toxin-evoked responses were blocked by amiloride (Amil, 1 mM). **c**, MitTx (75 nM)-evoked currents are comparable in magnitude to pH-4-evoked currents in ASIC1b-expressing oocytes. Toxin responses are non-desensitizing and persistent compared with transient proton-evoked currents. **d**, Dose–response analysis of toxin-evoked currents normalized to maximal pH-4-evoked response in ASIC-expressing oocytes. Data were fitted to the Hill equation. **e**, MitTx (75 nM) is a poor ASIC2a agonist, but dramatically potentiates pH-5.5-evoked responses. **f**, pH dose–response of ASIC2a in the absence (dark green) or presence (light green) of 75 nM MitTx. Data were fitted to the Hill equation. **g**, PcTx1 (100 nM) inhibits both pH-6- and MitTx-evoked currents in ASIC1a-expressing oocytes. **h**, MitTx occludes PcTx1 inhibition. Vertical scale bars, 1 μA; horizontal bars, 1 min; $V_h = -60$ mV.

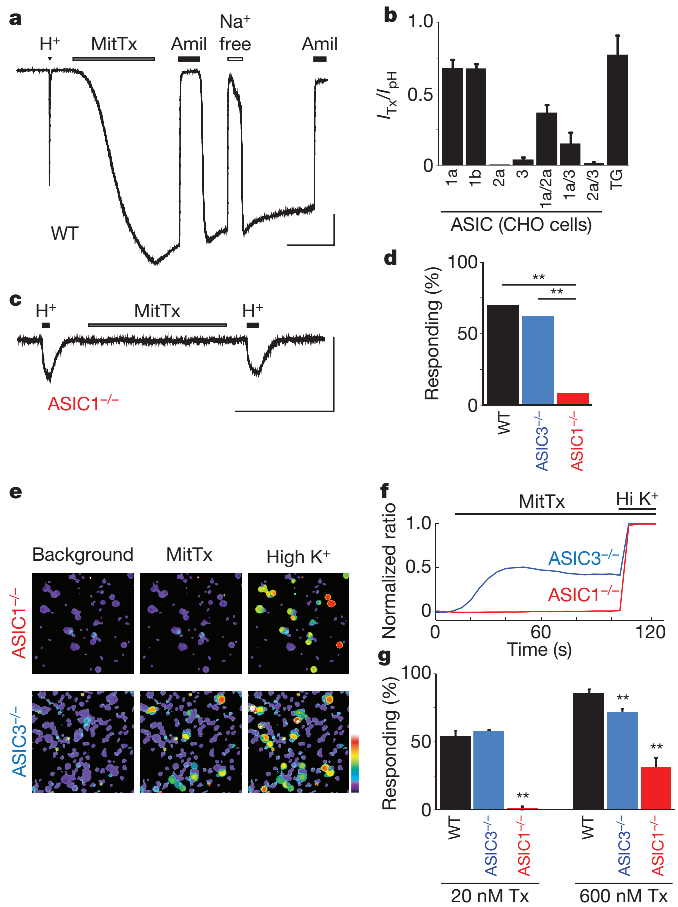


Figure 3 | ASICs are the neuronal receptor for MitTx. **a**, Whole-cell recording ($V_h = -60$ mV) from newborn rat trigeminal ganglia neuron shows representative pH 4 (H^+)- and MitTx (75 nM)-evoked responses. Toxin responses were blocked by amiloride (Amil; 1 mM), and eliminated when extracellular Na^+ was replaced with Cs^+ (Na^+ Free). **b**, MitTx (75 nM) activates homo- and heteromeric ASIC family members expressed in CHO cells. MitTx to pH current ratios for ASIC1a or 1b ($n = 3$–6) resembled profiles observed in trigeminal ganglia neurons ($n = 28$). pH 4 was used for all ratios except for measurements of ASIC1a, in which case pH 6 was used to minimize tachyphylaxis. **c**, Trigeminal ganglia neurons from newborn ASIC1$^{-/-}$ mice lacked MitTx sensitivity. H^+ indicates pH 4. **d**, Percentage of wild-type or knockout trigeminal ganglia neurons in which toxin-evoked currents were observed by whole-cell patch-clamp analysis ($n = 10$–30, $^{**}P < 0.01$, χ^2 test). **e**, MitTx (20 nM) activates trigeminal ganglia neurons from ASIC3$^{-/-}$, but not ASIC1$^{-/-}$, mice. **f**, Average MitTx-evoked calcium response of trigeminal ganglia neurons ($n > 300$) normalized to a high-potassium response (Hi K^+). **g**, Fraction of neurons responding to 20 nM or 600 nM MitTx assessed by calcium imaging ($n = 3$–4 trials, each with more than 100 cells; $^{**}P < 0.01$, one-way analysis of variance with *post hoc* Tukey's test). Vertical scale bars, 1 nA (**a**), 100 pA (**b**). Horizontal scale bars, 1 min. Error bars, mean ± s.e.m.

perfusate, and the relative magnitude of toxin-to-proton evoked responses resembled those observed in transfected mammalian (CHO) cells expressing the cloned rat ASIC1a or 1b channels (Fig. 3a, b). Further evidence that ASIC1 is the predominant target of MitTx came from analysis of electrophysiological responses in trigeminal neurons from newborn ASIC1- or ASIC3-deficient mice[19,20]. The percentage of toxin-sensitive cells from ASIC1$^{-/-}$ mice was greatly diminished compared with wild-type or ASIC3$^{-/-}$ animals (Fig. 3c, d). Using calcium imaging to sample a larger population of neurons, we found that responses to moderate toxin concentrations (20 nM) were completely absent in trigeminal neurons cultured from ASIC1$^{-/-}$ mice, but unperturbed in those from ASIC3$^{-/-}$ animals (Fig. 3e–g). Only at substantially higher toxin concentrations (600 nM; exceeding the EC_{50} for ASIC1 by at least 30-fold) did we observe a relatively small subset of toxin-sensitive neurons in ASIC1$^{-/-}$ cultures. These residual toxin responses were eliminated when depolarization-evoked calcium influx was suppressed in Na^+-free perfusate (Supplementary Fig. 6), consistent with Ca^{2+}-impermeant ASIC2 and/or ASIC3 subtypes accounting for the activity. Indeed, ASIC3$^{-/-}$ cultures showed a small but significant diminution in the percentage of neurons responding to 600 nM MitTx (Fig. 3g).

The role of ASIC channels in pain has focused primarily on ASIC3 given its somatosensory neuron-specific pattern of expression[13]. MitTx provides a novel tool with which to determine whether ASIC1 and ASIC1-expressing neurons also contribute to nociception.

Injection of MitTx into the hindpaw of wild-type mice produced robust nocifensive (pain-related) behaviour scored as a characteristic licking response (Fig. 4a). In the same animals, we observed abundant Fos protein expression in superficial laminae of the ipsilateral dorsal spinal cord, demonstrating engagement of nociceptive pathways (Fig. 4b). These responses were diminished in ASIC1-deficient mice, but persisted in ASIC3$^{-/-}$ animals, demonstrating a predominant contribution of ASIC1 channels to toxin-evoked nocifensive behaviour.

Are these responses mediated through a well-characterized population of nociceptors? To address this question, we first examined MitTx-sensitive neurons for expression of relevant molecular markers of subpopulations of nociceptors. One-third of toxin-sensitive neurons expressed TRPV1, whereas a much smaller group (11%) were labelled by isolectin B4, which in mice marks a population of TRPV1-negative, non-peptidergic C-fibres. Most (53%) toxin-sensitive neurons immunostained for the NF200 neurofilament, placing them among the subpopulation of medium-to-large diameter neurons with myelinated axons (Fig. 4c). As many large diameter sensory neurons respond to innocuous mechanical stimulation, these histological results suggest that functional ASIC1 channels are expressed by both nociceptive and non-nociceptive somatosensory neurons. Next, we asked whether MitTx elicits nocifensive responses in mice in which the central terminals of TRPV1-expressing nociceptors have been selectively ablated through spinal (intrathecal) injection of capsaicin[21]. Indeed, these animals showed a complete loss of toxin-evoked behaviour and Fos immunoreactivity in the spinal cord (Fig. 4d), demonstrating that the painful effects of MitTx are mediated entirely by a relatively small cohort (less than a third) of ASIC1-expressing nerve fibres.

Bites and stings from venomous creatures produce pain to ward off predators[2–4], and thus it stands to reason that some toxins have evolved to efficiently target elements of the pain pathway. In fact, bites from some coral snake species are well known to elicit excruciating pain requiring hospitalization and administration of opiate analgesics[9,22]. Our results show that MitTx has evolved from Kunitz-type and PLA2-like protein scaffolds to evoke intense and persistent pain by producing robust and long-lasting activation of ASIC1 channels on the nociceptive terminals of mammalian, avian or serpentine predators. Just as capsaicin and some spider toxins highlight the importance of TRPV1 in pain sensation, so MitTx implicates ASIC1 channels in this protective sensory modality, further illustrating the power of natural products in identifying key components and therapeutic targets of nociceptive signalling pathways.

In light of its selective expression within somatosensory ganglia and a well-documented contribution to ischaemic pain, studies of ASIC channels in nociception have focused primarily on the ASIC3 subtype[13,23–26]. The extremely persistent (non-desensitizing) and selective nature of MitTx action has enabled us to functionally isolate ASIC1 channels and reveal their contribution to pain sensation through the activation of TRPV1-expressing neurons. This nociceptor population has been suggested to constitute a 'labelled line' required for acute detection of noxious heat, as well as tissue injury-evoked pain hypersensitivity[1,21]. Activation of ASIC1 on these nerve fibres may contribute to the non-TRPV1 component of proton-mediated sensitization associated with tissue acidosis and inflammatory pain. ASIC1 is also expressed by non-nociceptive somatosensory neurons, as well as other neural and non-neural tissues[14,20,24,27]. Thus, MitTx represents a novel class of pharmacological probes with which to elucidate the contributions of ASIC channels to a variety of physiological processes. Finally, it is intriguing that MitTx can potentiate proton efficacy at some ASIC subtypes (most notably ASIC2a) by

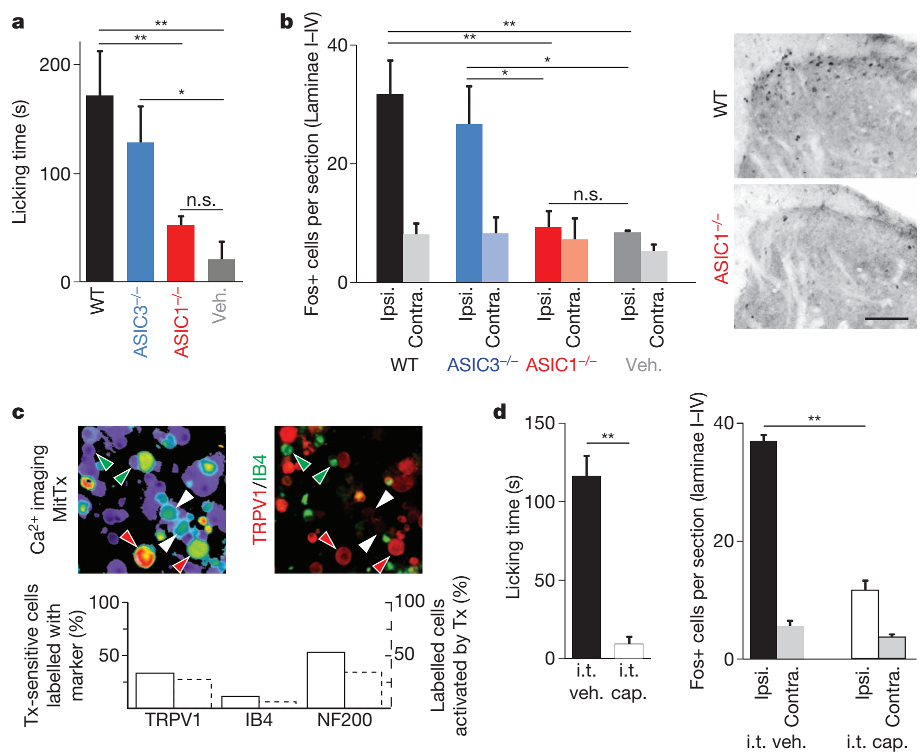


Figure 4 | MitTx elicits pain behaviour through ASIC1- and TRPV1-expressing nociceptors. **a**, Hindpaws of wild-type (WT) or ASIC-knockout mice were injected with MitTx (5 μM in 20 μl PBS with 0.1% BSA) or vehicle (Veh.) alone. Total time spent licking the injected paw was recorded over 15 min. **b**, Quantification and representative images of Fos immunostaining in superficial laminae of spinal cord sections from toxin-injected mice (ipsilateral (Ipsi.) or contralateral (Contra.) to the injection site). Scale bar, 50 μm. For **a** and **b**, $n = 4$–7; $*P < 0.05$; $**P < 0.01$, one-way analysis of variance with *post hoc* Tukey's test. **c**, Adult mouse DRG neurons were tested for toxin-sensitivity using calcium imaging, then stained with antibodies that mark specific subpopulations of cells. Red arrows, Tx-sensitive/TRPV1-positive cells; white arrows, Tx-sensitive/TRPV1-negative cells; green arrows, IB4-positive/Tx-insensitive cells. Percentages were counted for more than 200 cells per marker and graphed below. **d**, Intrathecal administration of capsaicin (i.t. cap., 10 μg in 5 μl), but not vehicle (i.t. veh.), eliminates behavioural response and spinal Fos induction after intraplantar toxin injection (as in **a**, **b**); $n = 3$–4; $**P < 0.001$, Student's *t*-test. Error bars, mean ± s.e.m.

two or three orders of magnitude. That is to say, MitTx reveals the fact that protons only activate ASIC2a channels to less than 10% of maximal open probability, suggesting that protons do not exploit the full potential of ASIC2a as a ligand-gated channel. Although small FMRF-amide-like peptides have been shown to potentiate proton-gated currents through ASIC1 and ASIC3 (primarily by slowing desensitization)[25,28], the profound enhancing effects mediated by MitTx hint at the existence of other, more potent physiological modulators for this class of excitatory channels.

METHODS SUMMARY

MitTx-α and MitTx-β were purified from crude *M. tener tener* venom (National Natural Toxin Research Center, Texas A&M University-Kingsville) using multiple reversed-phase high-performance liquid chromatography (HPLC) steps and purity assessed by mass spectrometry and gel filtration chromatography. Partial toxin sequences were determined by Edman degradation and/or *de novo* tandem mass spectrometry sequencing analysis and used to design degenerate primers for cloning full-length MitTx-α and MitTx-β cDNAs from *M. t. tener* venom gland library. PcTx1 was purified from *Psalmopoeus cambridgei* venom (SpiderPharm) and *M. frontalis* venom was purchased from Sigma. ASICs were cloned from rat trigeminal ganglion or brain cDNA libraries (except mouse ASIC2b, which was provided by M. Welsh). For isothermal titration calorimetry (ITC) experiments, MitTx-β (3–10 μM) was titrated into MitTx-α (0.3–1 μM) at 15 °C. Molecular biology, neuronal dissociation, cell culture, calcium imaging and electrophysiological experiments were conducted essentially as described[6]. P_{Na^+}/P_{Cs^+} was determined by substituting observed reversal potential into the Goldman–Hodgkin–Katz voltage equation. Animal experiments were approved by the University of California, San Francisco, (UCSF) Institutional Animal Care and Use Committee and conducted in accordance with the National Institutes of Health (NIH) Guide for the Care and Use of Laboratory Animals and the recommendations of the International Association for the Study of Pain as described[5,21]. Behavioural and histological experiments were both conducted and scored with the experimenter blind to genotype.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 August; accepted 15 October 2011.

1. Basbaum, A. I., Bautista, D. M., Scherrer, G. & Julius, D. Cellular and molecular mechanisms of pain. *Cell* **139,** 267–284 (2009).
2. Mebs, D. *Venomous and Poisonous Animals: A Handbook for Biologists, and Toxicologists and Toxinologists, Physicians and Pharmacists* (CRC Press, 2002).
3. Chahl, L. A. & Kirk, E. J. Toxins which produce pain. *Pain* **1,** 3–49 (1975).
4. Schmidt, J. O. Biochemistry of insect venoms. *Annu. Rev. Entomol.* **27,** 339–368 (1982).
5. Siemens, J. *et al.* Spider toxins activate the capsaicin receptor to produce inflammatory pain. *Nature* **444,** 208–212 (2006).
6. Bohlen, C. J. *et al.* A bivalent tarantula toxin activates the capsaicin receptor, TRPV1, by targeting the outer pore domain. *Cell* **141,** 834–845 (2010).
7. Terlau, H. & Olivera, B. M. Conus venoms: a rich source of novel ion channel-targeted peptides. *Physiol. Rev.* **84,** 41–68 (2004).
8. Fry, B. G. *et al.* The toxicogenomic multiverse: convergent recruitment of proteins into animal venoms. *Annu. Rev. Genomics Hum. Genet.* **10,** 483–511 (2009).
9. Morgan, D. L., Borys, D. J., Stanford, R., Kjar, D. & Tobleman, W. Texas coral snake (*Micrurus tener*) bites. *South. Med. J.* **100,** 152–156 (2007).
10. Doley, R. & Kini, R. M. Protein complexes in snake venom. *Cell. Mol. Life Sci.* **66,** 2851–2871 (2009).
11. Berg, O. G., Gelb, M. H., Tsai, M. D. & Jain, M. K. Interfacial enzymology: the secreted phospholipase A(2)-paradigm. *Chem. Rev.* **101,** 2613–2654 (2001).
12. Waldmann, R., Champigny, G., Bassilana, F., Heurteaux, C. & Lazdunski, M. A proton-gated cation channel involved in acid-sensing. *Nature* **386,** 173–177 (1997).
13. Waldmann, R. *et al.* Molecular cloning of a non-inactivating proton-gated Na^+ channel specific for sensory neurons. *J. Biol. Chem.* **272,** 20975–20978 (1997).
14. Wu, L. J. *et al.* Characterization of acid-sensing ion channels in dorsal horn neurons of rat spinal cord. *J. Biol. Chem.* **279,** 43716–43724 (2004).
15. Escoubas, P. *et al.* Isolation of a tarantula toxin specific for a class of proton-gated Na^+ channels. *J. Biol. Chem.* **275,** 25116–25121 (2000).
16. Chen, X., Kalbacher, H. & Grunder, S. Interaction of acid-sensing ion channel (ASIC) 1 with the tarantula toxin psalmotoxin 1 is state dependent. *J. Gen. Physiol.* **127,** 267–276 (2006).
17. Leffler, A., Monter, B. & Koltzenburg, M. The role of the capsaicin receptor TRPV1 and acid-sensing ion channels (ASICS) in proton sensitivity of subpopulations of primary nociceptive neurons in rats and mice. *Neuroscience* **139,** 699–709 (2006).
18. Poirot, O., Berta, T., Decosterd, I. & Kellenberger, S. Distinct ASIC currents are expressed in rat putative nociceptors and are modulated by nerve injury. *J. Physiol. (Lond.)* **576,** 215–234 (2006).
19. Price, M. P. *et al.* The DRASIC cation channel contributes to the detection of cutaneous touch and acid stimuli in mice. *Neuron* **32,** 1071–1083 (2001).
20. Wemmie, J. A. *et al.* The acid-activated ion channel ASIC contributes to synaptic plasticity, learning, and memory. *Neuron* **34,** 463–477 (2002).
21. Cavanaugh, D. J. *et al.* Distinct subsets of unmyelinated primary sensory fibers mediate behavioral responses to noxious thermal and mechanical stimuli. *Proc. Natl Acad. Sci. USA* **106,** 9075–9080 (2009).
22. Nishioka, S. A., Silveira, P. V. & Menzes, L. B. Coral snake bite and severe local pain. *Ann. Trop. Med. Parasitol.* **87,** 429–431 (1993).
23. Sutherland, S. P., Benson, C. J., Adelman, J. P. & McCleskey, E. W. Acid-sensing ion channel 3 matches the acid-gated current in cardiac ischemia-sensing neurons. *Proc. Natl Acad. Sci. USA* **98,** 711–716 (2001).
24. Wemmie, J. A., Price, M. P. & Welsh, M. J. Acid-sensing ion channels: advances, questions and therapeutic opportunities. *Trends Neurosci.* **29,** 578–586 (2006).
25. Deval, E. *et al.* Acid-sensing ion channels (ASICs): pharmacology and implication in pain. *Pharmacol. Ther.* **128,** 549–558 (2010).
26. Yu, Y. *et al.* A nonproton ligand sensor in the acid-sensing ion channel. *Neuron* **68,** 61–72 (2010).
27. Ziemann, A. E. *et al.* The amygdala is a chemosensor that detects carbon dioxide and acidosis to elicit fear behavior. *Cell* **139,** 1012–1021 (2009).
28. Askwith, C. C. *et al.* Neuropeptide FF and FMRFamide potentiate acid-evoked currents from sensory neurons and proton-gated DEG/ENaC channels. *Neuron* **26,** 133–141 (2000).
29. Jacobson, M. P. *et al.* A hierarchical approach to all-atom protein loop prediction. *Proteins* **55,** 351–367 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Price and M. Welsh for providing ASIC1 and ASIC3 knockout mice; Y. Kelly and J. Poblete for technical assistance; C. Williams for assisting with homology models; F. Findeisen, L. Ma and D. Minor for assistance with ITC experiments; R. Nicoll and members of the Julius laboratory for discussion and comments. This work was supported by a Ruth Kirschstein NIH predoctoral fellowship (F31NS065597 to C.B.), an NIH postdoctoral training grant from the UCSF Cardiovascular Research Institute (to A.C.), a postdoctoral fellowship from the Canadian Institutes of Health Research (to R.S.-N.), the Howard Hughes Medical Institute (K.F.M. and A.L.B.), and the NIH (NCRR P41RR001614 to A.L.B., NCRR P40RR018300-09 to E.S. and NINDS R01NS065071 to D.J.).

Author Contributions C.B. and A.C. initiated the screen, performed experiments and analysed data. R.S.-N. and A.I.B. performed and analysed behavioural experiments and spinal cord histology. K.F.M., A.L.B., S.Z. and D.K. determined partial protein sequences and performed mass spectrometry measurements. E.S. provided snake venom and tissue. C.B., A.C. and D.J. wrote the manuscript with discussion and contribution from all authors. D.J. supervised the project and provided guidance throughout.

Author Information MitTx-α, MitTx-β and MttPLA2 cDNA sequences are deposited in GenBank under accession numbers JN613325, JN613326 and JN613327, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.J. (david.julius@ucsf.edu).

METHODS

Toxin purification. Crude *M. tener tener* venom, pooled from multiple specimens, was provided by the National Natural Toxins Research Center, Texas A&M University-Kingsville, Texas, USA. Lyophilized venom was dissolved in water to $100\,\mathrm{mg\,ml^{-1}}$, diluted 30-fold into 20% acetonitrile containing 0.1% trifluoroacetic acid (TFA), filtered through a 0.1-μm centrifugal filter unit (Millipore) and fractionated by reversed-phase HPLC. Up to 20 mg venom was loaded onto a semi-preparative C18 column (Vydac model 218TP510), and eluted with a 20-min linear gradient (18–36% acetonitrile; $3\,\mathrm{ml\,min^{-1}}$). Fractions containing predominantly MitTx-α were diluted twofold with 0.1% TFA, injected onto an analytical PLRP-S column (Varian PL1512-5501) and separated with an 8-min linear gradient (27–34% acetonitrile; $0.8\,\mathrm{ml\,min^{-1}}$); MitTx-α-containing fractions were again diluted, injected onto an analytical C18 column (Vydac 218TP54) and separated with a 20-min linear gradient (18–36% acetonitrile; $0.8\,\mathrm{ml\,min^{-1}}$). MitTx-β fractions from the semi-preparative run were diluted, applied to an analytical C18 column (Vydac 218TP54) and separated with a 7-min linear gradient (30–36% acetonitrile; $0.8\,\mathrm{ml\,min^{-1}}$). All HPLC buffers contained 0.1% TFA, and all purifications were performed at room temperature. Purified fractions were lyophilized then dissolved in water, aliquoted, and stored at −80 °C. Protein concentration was determined using calculated extinction coefficient at 280 nm (http://us.expasy.org /tools/protparam.html).

PcTx1 was purified from *Psalmopoeus cambridgei* venom (SpiderPharm) using two sequential HPLC steps consisting of 114-min linear gradient (0–54% acetonitrile on a semi-preparative C18 column) followed by the same gradient on an analytical C18 column (as above). *M. frontalis* venom was purchased from Sigma.

Sequence determination. Toxin masses were determined using a Bruker Apex III ESI-Q-FTICR mass spectrometer, and the N-terminal sequencing was performed on an Applied Biosystems 492 Procise Sequencer previously described[6]. Edman degradation yielded partial sequence (NLNQFRLMIKCTNDRV…) for MitTx-β, but not MitTx-α owing to modified (pyroglutamic acid) N terminus. Partial MitTx-α sequence was determined *de novo* by tandem mass spectrometry sequencing analysis. After reduction under acidic conditions (0.1% formic acid, 5 mM TCEP for 24 h at 37 °C) MitTx-α was subjected to collision-induced dissociation (CID), higher-energy collisional dissociation (HCD) and electron transfer dissociation (ETD) analyses using an LTQ-Orbitrap mass spectrometer (Thermo Fisher). Data were analysed manually, yielding the N-terminal sequence as well as a shorter internal sequence fragment (<Q[L/I]RPAFCYEDPPFFQKCGAFVDSYYF... and …HFFYGQCDV…).

Partial protein sequences were used to clone full-length MitTx-α and MitTx-β cDNAs. RNA was extracted from two *M. t. tener* venom glands using TRIZOL reagent (Invitrogen), then isolated by chloroform extraction and isopropanol precipitation. A cDNA library was generated for use in 5′ and 3′ rapid amplification of cDNA ends (RACE) reactions using SMARTer RACE cDNA Amplification Kit (Clontech). Primers derived from biochemically determined sequences and, for MitTx-β, taking advantage of conservation in reported *Micrurus* PLA2 sequences, small fragments of MitTx-α and MitTx-β sequences were amplified by PCR and then used to design gene-specific 3′ and 5′ RACE primers. RACE products were sequenced individually after insertion into TOPO vector (Invitrogen), and each sequence fragment was confirmed by multiple sequencing reads and multiple RACE primer sets. The cDNA-derived peptide sequence predicted the observed molecular weight of the purified toxins after consideration of the post-translational modifications (N-terminal cyclization and disulphide bond formation).

Calorimetry. MitTx-α and MitTx-β were diluted into ITC buffer (150 mM NaCl, 1 mM $CaCl_2$, 10 mM HEPES, adjusted to pH 7.4 with NaOH). Toxin aliquots were diluted with water before dilution with ITC buffer to maintain the same dilution factor. Diluted toxin solutions were centrifuged (65,000*g* for 30 min at 4 °C), degassed (5 min at 15 °C), and loaded into a VP-ITC Microcalorimeter (MicroCal). MitTx-β (3–10 μM) was titrated into MitTx-α (0.3–1 μM) following a schedule of one 4-μl injection followed by 29 injections of 10 μl, each spaced by 5 min. Titrations were conducted at 15 °C. After manual baseline correction, total heat released per injection was calculated by integrating over the full injection time period. Heat of dilution was estimated by the final titration points and subtracted from baseline. Linear baseline values were confirmed by titration of MitTx-β into buffer. Data were processed and fit to a single-site binding model using Origin version 7 (MicroCal). Reported values are averages of three independent experiments.

Molecular biology. Full-length (not including untranslated regions) ASIC1a, 1b, 2a, 3 and 4 were cloned from rat trigeminal ganglion or brain cDNA libraries into pCDNA3.1 (Invitrogen). The mouse ASIC2b clone was provided by M. Welsh, and rat ENaC clones were provided by D. Pearce. All constructs were confirmed by DNA sequencing. Other constructs and RNA synthesis have been previously described[6].

***Xenopus* oocyte and CHO/HEK cell culture.** *Xenopus laevis* oocytes (Nasco), human embryo kidney 293 (HEK293) and Chinese hamster ovary (CHO-K1) cells were isolated, maintained, transfected and plated essentially as described[6]. Oocytes were injected with RNA (3–50 ng; 50 nl) and assayed 3–7 days later. CHO culture medium included non-essential amino acids (UCSF Cell Culture Facility). Electrophysiological experiments were conducted 2–20 h after plating, and imaging experiments were performed 3–4 h after plating.

Neuronal cell culture. Trigeminal ganglia were dissected from newborn (P0–P2) Sprague-Dawley rats or C57BL/6 mice and cultured as described[6] for 3–4 h before calcium imaging or 2–20 h before electrophysiological recordings. Neurons dissected from newborn animals were used for all experiments except histology, which used adult neurons owing to apparent developmental changes in ASIC subtype expression (see Supplementary Fig. 4c). Dorsal root ganglia from adult (4- to 12-week-old) mice were dissected and dissociated as newborn neurons, except collagenase P and trypsin incubations were extended to 30 min each. After trituration, cells were centrifuged through a 15% BSA gradient for 5 min at 1,000*g* to remove cellular debris. BSA was rinsed from the pellet and cells were re-suspended in culture medium. Cells were plated into PDL-coated 384-well plates (Greiner Bio-One) for 14–18 h before calcium imaging.

Calcium imaging. Cells were loaded with Fura-2-AM as described[6]. MitTx-α, MitTx-β and PcTx1 solutions were prepared in the presence of 0.1% bovine serum albumin (BSA, Sigma) to minimize toxin adsorption to plasticware. For experiments using low concentrations of MitTx-α/β, MitTx-α and MitTx-β were first mixed at high concentrations (≥1 μM) for at least 10 min before dilution.

Electrophysiology. *Xenopus* oocyte two-electrode and mammalian whole-cell recordings were performed essentially as described[6]. Oocyte extracellular solution contained 115 mM NaCl, 2.5 mM KCl, 1.8 mM $MgCl_2$, 5 mM HEPES and 5 mM MES adjusted to pH 5–7.4 with NaOH. For pH < 5, citrate was used instead of HEPES/MES. Solutions were applied using gravity-based perfusion over a small-volume oocyte chamber (Automate Scientific), except for toxin-containing solutions, which were pipetted directly onto cells.

Mammalian cell extracellular solution contained 150 mM NaCl, 2.8 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 5 mM HEPES, 5 mM MES adjusted to pH 7.4 with NaOH, 300–310 $\mathrm{mOsmol\,kg^{-1}}$. The pipette solution contained 130 mM K-gluconate, 15 mM KCl, 4 mM NaCl, 0.5 mM $CaCl_2$, 1 mM EGTA, 10 mM HEPES, adjusted to pH 7.2 with KOH, 285 $\mathrm{mOsmol\,kg^{-1}}$. Solutions were applied from the micro-perfusion system SmartSquirt (Automate Scientific). Toxins were applied in the presence of 0.1% BSA and MitTx-α/β complex was formed at high concentrations as described above.

Fits to the Hill equation were performed using Igor Pro software (Wavemetrics) with four free parameters. Fits in Fig. 2d revealed the following values for EC_{50}, n_H and maximum respectively: 9.4 ± 1.3 nM, 2.4 ± 0.8 and 1.30 ± 0.08 for ASIC1a; 23 ± 1 nM, 3.6 ± 0.7 and 1.35 ± 0.05 for ASIC1b; 36 ± 4 nM, 0.088 ± 0.005 and 2.2 ± 0.5 for ASIC2a; and 830 ± 250 nM, 1.4 ± 0.3 and 1.8 ± 0.3 for ASIC3. Fits in Fig. 2f revealed the following values for pH_{50}, n_H and maximum respectively: pH 5.98 ± 0.07, 1.3 ± 0.3 and 0.99 ± 0.03 in the presence of toxin and pH 2.4 ± 0.1, 1.0 ± 0.1 and 1.6 ± 4.1 in the absence of toxin. The baseline parameter was near-zero for all fits.

For permeability experiments from acutely dissociated rat trigeminal ganglia neurons, endogenous currents were attenuated by using a minimal ionic composition and by applying rapid voltage ramps (140-ms ramps from −80 mV to +80 mV were applied every 200 ms) to drive inactivating voltage-gated channels into a non-conductive state; extracellular and intracellular solutions contained 150 mM NaCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 10 mM HEPES adjusted to pH 7.4 with NaOH. Remaining endogenous currents were negligible in magnitude compared with the very large MitTx-evoked currents, so no baseline subtraction was applied. Dialysis of the intracellular solution was monitored to completion as determined by elimination of K_v channel currents. Caesium wash solution contained 150 mM CsCl instead of NaCl, and pH was titrated with CsOH. P_{Na+}/P_{Cs+} was calculated by substituting the observed reversal potential (V_m) into the Goldman–Hodgkin–Katz voltage equation: $P_{Na+}/P_{Cs+} = e^{-Vm(F/RT)}$ where F, R and T have their usual meanings ($F/RT = 39.6\,\mathrm{V^{-1}}$ at 20 °C). Currents were not corrected for liquid junction potential changes.

Immunohistochemistry. After calcium imaging, adult DRG neurons were fixed with 4% paraformaldehyde (10 min) then washed three times in PBS containing 0.1% triton X (PBSTx). Neurons were incubated in blocking solution for 1 h (10% normal goat serum (NGS) + PBSTx) and then incubated overnight in primary antibody solution (2.5% NGS + PBSTx + primary antibody; rabbit anti-TRPV1, 1:8,000 or mouse anti-N5A; Sigma, 1:1,000). Cells were washed three times with PBSTx before a incubation for 2 h in secondary antibody solution (2.5% NGS + PBSTx, + Alexa-488 or Alexa-594; Invitrogen, 1:1,000). For IB4 staining, IB4-FITC (Sigma, $10\,\mathrm{\mu g\,ml^{-1}}$) was added to the secondary antibody solution. Cells were washed three times in PBS before imaging.

For spinal cord Fos immunoreactivity, mice were perfused transcardially with 4% paraformaldehyde 90 min after hindpaw toxin injection (as for behavioural experiments). Frozen spinal cord sections (25 μm) were prepared from lumbar level L4/L5 and immunostained for Fos as described[30].

Behaviour. Animal experiments were approved by the UCSF Institutional Animal Care and Use Committee and conducted in accordance with the NIH Guide for the Care and Use of Laboratory Animals and the recommendations of the International Association for the Study of Pain. Two to five animals were housed per cage and maintained on a 12-h light/dark schedule with ad libitum access to food and water. Injections (20 μl PBS + 0.1% BSA, with or without 5 μM MitTx) were performed as described on adult (10- to 18-week-old), male mice[5]. For intrathecal capsaicin studies, adult mice were anaesthetized and treated as described[21]. Behavioural tests were performed 1–5 days after capsaicin injection. Knockout strains were extensively backcrossed to the C57BL/6 wild-type strain. Behavioural and histological experiments were both conducted and scored with the experimenter blind to genotype.

Primer sequences. MitTx-β fragment forward: AACCTCTAYCAGTTCATGATTAAATGTACCAACG; MitTx-β fragment reverse: TCATTGGCAACGTTTGAGGTCGATATTG; MitTx-β 3′ RACE: GTTATCTAGCCAGCGACCTCGATTGCAGTGG; MitTx-β 5′ RACE: CCACTGCAATCGAGGTCGCTGGCTAGATAAC; MitTx-α fragment forward: CCNCCNTTYTTYCARAARTGYGGNGCNTTYGTNG; MitTx-α fragment reverse: ACRTCRCAYTGNCCRTARAARAARTG; MitTx-α 3′ RACE: CCTCCATTCTTTCAAAAATGTGGAGCC; MitTx-α 5′ RACE: CGCAAGTAATTCTTGACCTGTTGAAGTAGTAGG.

30. Caterina, M. J. *et al.* Impaired nociception and pain sensation in mice lacking the capsaicin receptor. *Science* **288,** 306–313 (2000).

LETTER

doi:10.1038/nature10534

Homeostatic response to hypoxia is regulated by the N-end rule pathway in plants

Daniel J. Gibbs[1]*, Seung Cho Lee[2]*, Nurulhikma Md Isa[1], Silvia Gramuglia[1], Takeshi Fukao[2], George W. Bassel[1], Cristina Sousa Correia[1], Françoise Corbineau[3], Frederica L. Theodoulou[4], Julia Bailey-Serres[2] & Michael J. Holdsworth[1]

Plants and animals are obligate aerobes, requiring oxygen for mitochondrial respiration and energy production. In plants, an unanticipated decline in oxygen availability (hypoxia), as caused by roots becoming waterlogged or foliage submergence, triggers changes in gene transcription and messenger RNA translation that promote anaerobic metabolism and thus sustain substrate-level ATP production[1]. In contrast to animals[2], oxygen sensing has not been ascribed to a mechanism of gene regulation in response to oxygen deprivation in plants. Here we show that the N-end rule pathway of targeted proteolysis acts as a homeostatic sensor of severe low oxygen levels in *Arabidopsis*, through its regulation of key hypoxia-response transcription factors. We found that plants lacking components of the N-end rule pathway constitutively express core hypoxia-response genes and are more tolerant of hypoxic stress. We identify the hypoxia-associated ethylene response factor group VII transcription factors of *Arabidopsis* as substrates of this pathway. Regulation of these proteins by the N-end rule pathway occurs through a characteristic conserved motif at the amino terminus initiating with Met-Cys. Enhanced stability of one of these proteins, HRE2, under low oxygen conditions improves hypoxia survival and reveals a molecular mechanism for oxygen sensing in plants via the evolutionarily conserved N-end rule pathway. SUB1A-1, a major determinant of submergence tolerance in rice[3], was shown not to be a substrate for the N-end rule pathway despite containing the N-terminal motif, indicating that it is uncoupled from N-end rule pathway regulation, and that enhanced stability may relate to the superior tolerance of Sub1 rice varieties to multiple abiotic stresses[4].

The N-end rule pathway of targeted proteolysis associates the fate of a protein substrate with the identity of its N terminus (the N degron)[5,6]. The N-terminal residue is classified as stabilizing or destabilizing, depending on the fate of the protein. An N degron containing a destabilizing residue is created through specific proteolytic cleavage, but can also be generated via successive enzymatic or chemical modifications to the N terminus; for example, arginylation by Arg-tRNA protein transferases (ATE)[7–9] (Supplementary Fig. 1). N-end rule pathway substrates containing destabilizing residues are targeted for proteasomal degradation via specific E3 ligases (also known as N recognins), such as PROTEOLYSIS1 and PROTEOLYSIS6 (PRT1 and PRT6) in *Arabidopsis*, which accept substrates with hydrophobic and basic N termini, respectively[8–10]. Several substrates of the N-end rule pathway are important developmental regulators in mammals[11] but as yet no substrates have been identified in plants. Previously we showed a function of this pathway in abscisic acid (ABA) signalling through PRT6 and ATE[12], and it has also been associated with leaf senescence and shoot and leaf development in *Arabidopsis*[13,14]. To understand N-end-rule-pathway-regulated gene expression we analysed the transcriptome of imbibed seed and seedlings of N-end rule pathway mutants *ate1 ate2*, which lack ATE activity[14], and *prt6* (Fig. 1a and Supplementary Table 1). This analysis revealed that genes important for anaerobic metabolism and survival of hypoxia, such as *ADH1*, *SUS4* and *PDC1*, were constitutively expressed at high levels in both mutants, in common with wild-type Col-0 plants under hypoxia (Supplementary Fig. 2). For example, 47 of the 135 differentially regulated mRNAs in the wild-type hypoxia-induced transcriptome were also upregulated in *prt6* seedlings grown under non-stress conditions (Supplementary Table 1; signal $\log_2$ ratio ≥ 1, false discovery rate ≤ 0.01). The mRNAs upregulated in *prt6* and *ate1 ate2* mutants included over half of the core 49 mRNAs upregulated by hypoxia across seedling cell types[15] (Fig. 1b and Supplementary Fig. 2). Consistent with this observation, β-glucuronidase (GUS) expression driven by the promoter of *ADH1* (*pADH1::GUS*; ref. 16) was upregulated in wild-type seedlings subjected to hypoxia and ectopically expressed in mature embryos, roots and lower hypocotyls of *prt6* mutants (Fig. 1c and Supplementary Fig. 3). Constitutive expression of hypoxia-induced genes by N-end rule pathway mutant seedlings suggested that they would be resistant to hypoxic conditions. Imbibed seeds of both *prt6* and *ate1 ate2* mutants were able to germinate well under low oxygen (3%) compared to wild type (Fig. 1d), and mutant seedlings were more able to survive prolonged oxygen deprivation (Fig. 1e, f). The *ate1 ate2* double mutant showed greater resistance to hypoxia than *prt6*, indicating the existence of other as-yet-unidentified Arg-related E3 ligases, as previously postulated[10,14].

Transcription factors of the five-member *Arabidopsis* ethylene response factor (ERF) group VII[17] have recently been shown to enhance plant responses to hypoxia or anoxia, including HYPOXIA RESPONSIVE1 and 2 (HRE1 and HRE2)[18] and RELATED TO AP2 2 (RAP2.2)[19]. Overexpression of RAP2.12 was also shown to induce expression of a *pADH1::LUCIFERASE* reporter gene[20]. This subfamily shows homology to the agronomically important rice ERFs SUBMERGENCE 1A, B and C (ref. 3) and SNORKEL 1 and 2 (ref. 21). *SUB1A-1* within the *SUBMERGENCE 1* (*SUB1*) locus (which also contains *SUB1B* and *SUB1C*) was shown to be a primary determinant of enhanced survival of rice plants under complete submergence[3]. With the exception of SUB1C, all contain the initiating motif Met-Cys (MC) at the N terminus, embedded within a longer consensus shared with most other group VII ERFs of *Arabidopsis* and rice, MCGGAII (Supplementary Fig. 4a).

Removal of N-terminal methionine by METHIONINE AMINOPEPTIDASE (MAP) reveals the tertiary destabilizing residue cysteine in proteins initiating with MC, which targets substrates for degradation by the N-end rule pathway[7,9,22] (Supplementary Fig. 1). In mouse, N-end-rule-pathway-mediated degradation of the MC-motif-containing G-protein signalling components RGS4 and RGS5 is perturbed under hypoxia[22,23]. It was hypothesized that oxidation of cysteine at position 2 (C2) in these proteins under normoxia creates a secondary destabilizing

[1]Division of Plant and Crop Sciences, School of Biosciences and Centre for Plant Integrative Biology, University of Nottingham, Loughborough LE12 5RD, UK. [2]Center for Plant Cell Biology and Department of Botany and Plant Sciences, University of California, Riverside, California 92521, USA. [3]UPMC Univ Paris 06, UR5-EAC 7180 CNRS, Boîte courrier 156, 4 place Jussieu, F-75005 Paris, France. [4]Biological Chemistry Department, Rothamsted Research, Harpenden AL5 2JQ, UK.
*These authors contributed equally to this work.

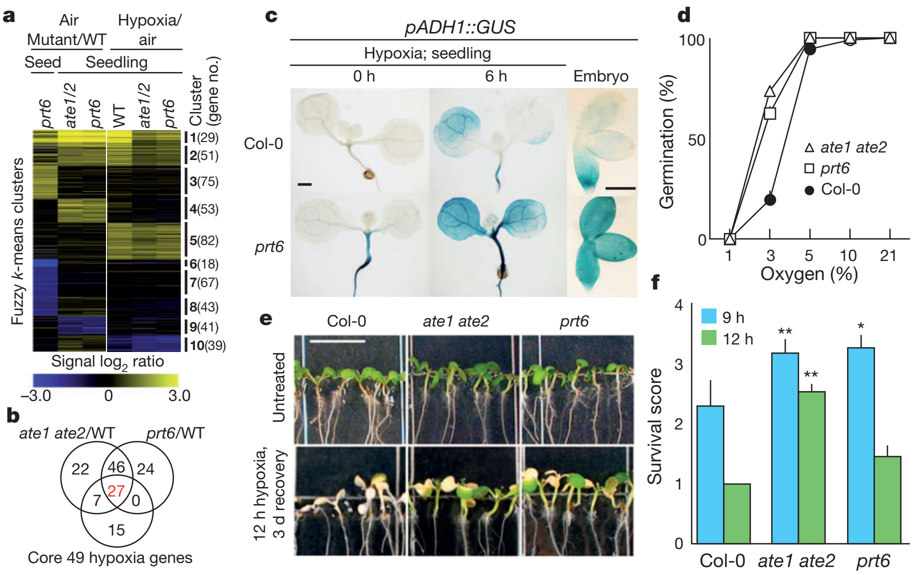


Figure 1 | N-end rule mutants ectopically accumulate anaerobic response mRNAs and are more tolerant to hypoxia. **a**, Expression data for differentially expressed genes comparing wild-type (Col-0) and mutants under air or hypoxia (2 h $-O_2$). **b**, mRNAs upregulated in mutants overlap with 49 mRNAs induced across cell types by hypoxia in wild-type seedlings[15]. **c**, Spatial visualization of *ADH1* promoter activity. Scale bars: 100 μm. **d**, Germination under reduced oxygen availability. **e**, Seedlings after 12 h of hypoxia and 3 d recovery. Scale bar: 0.6 cm. **f**, N-end rule pathway mutants are less sensitive to hypoxia stress. Data are mean of replicate experiments ± s.d.; $^*P < 0.05$; $^{**}P < 0.01$.

residue allowing addition of arginine (R) to the N terminus by ATE, creating a primary destabilizing residue[23]. We investigated the possibility that all *Arabidopsis* group VII ERFs as well as rice SUB1A-1 are N-end rule pathway substrates. A heterologous rabbit reticulocyte lysate assay[23] was used to express haemagglutinin (HA)-tagged ERFs driven by a T7 promoter *in vitro*, because components of the N-end rule pathway (ATE, MAP and PRT6) are highly conserved in eukaryotes[8], and it has been shown that wheat-germ lysate does not contain an active proteosomal system[24]. *Arabidopsis* group VII ERFs were short-lived, and their stability was enhanced by MG132 and the N-end rule pathway competitive dipeptide Arg-β-Ala, but not by the non-competitive Ala-Ala dipeptide[23] (Fig. 2a). Mutation of C2 to alanine (C2A), which should remove the N-degron and stabilize proteins specifically with respect to the N-end rule pathway[23], significantly enhanced stability *in vitro* of *Arabidopsis* ERFs, indicating that all group VII ERFs are potential substrates of the N-end rule pathway. *Arabidopsis* contains 206 proteins from gene models with MC at the N terminus; we used two of these—VERNALISATION 2 (VRN2) and MADS AFFECTING FLOWERING 5 (MAF5), which lack the extended N-terminal group VII ERF consensus (Supplementary Fig. 4b)—to test the specificity of this sequence. Whereas HA-tagged VRN2 (VRN2–HA) was degraded in this system, and stabilized by the introduction of a C2A mutation (VRN2(C2A)–HA), MAF5–HA and MAF5(C2A)–HA were both stable (Fig. 2b), indicating that not all *Arabidopsis* MC proteins are N-end rule pathway substrates. This is not surprising as it has previously been shown that optimal positioning of a downstream lysine for ubiquitination is also a key determinant of the quality of an N degron[8,9,25]. SUB1A-1 was resistant to degradation (Fig. 2c). As the N-terminal sequence of SUB1A-1 differs at position 5 (E rather than A, Supplementary Fig. 4a), we analysed a mutant version that replaced this amino acid to reconstitute the consensus group VII sequence (SUB1A-1(E5A)–HA). SUB1A-1(E5A)–HA was also stable *in vitro* (Fig. 2c), indicating that degradation of this protein is uncoupled from the N-end rule pathway. As expected, the rice protein SUB1C-1–HA, lacking an MC N terminus, was long lived *in vitro* (Fig. 2c).

To confirm the activity of the N-end rule pathway towards specific MC-containing substrates in plants, we analysed the *in vivo* longevity of the ERF proteins HRE1 and HRE2 (Fig. 2d). We expressed either wild-type or mutant (HRE1(C2A), HRE2(C2A)) HA-tagged versions of these proteins ectopically using the CaMV35S promoter in *Arabidopsis*. In wild-type plants, only the mutant C2A proteins could be detected at high levels, despite detectable expression of corresponding mRNAs, indicating that wild-type versions are N-end rule pathway substrates *in vivo*. HRE2–HA expressed in the *prt6* mutant was stable, linking its degradation directly to PRT6. To assess whether oxygen regulates the stability of HRE proteins, we analysed the accumulation of HRE–HA proteins in wild-type plants expressing HRE1–HA, HRE1(C2A)–HA, HRE2–HA and HRE2(C2A)–HA under normal and low oxygen conditions (Fig. 3a). After transfer of seedlings to hypoxic conditions we observed elevation of HRE2–HA within 2 h, but could not detect HRE1–HA (Fig. 3a and Supplementary Fig. 5a, b). HRE2–HA became destabilized again upon return to normoxic conditions (Fig. 3a). Both seeds and seedlings ectopically expressing stable C2A versions of HRE1 and HRE2 had increased tolerance to extended periods of oxygen deprivation (Fig. 3b–d and Supplementary Fig. 5c).

These data demonstrate that *Arabidopsis* ERF group VII transcription factors are substrates of the N-end rule pathway, and function to sense molecular oxygen, most likely through oxidation of the tertiary destabilizing residue cysteine. Stabilization of these proteins under hypoxic conditions leads to increased survival under low oxygen stress (Fig. 3e). It is currently unclear whether oxidation occurs through a chemical or enzymatic mechanism, although cysteine is readily oxidized chemically[26]. It is also unclear whether oxidation is related directly to molecular oxygen, or if indirect cellular changes associated with oxygen availability (such as alterations in cytosolic pH[27] and specific metabolites or transient accumulation of reactive oxygen species[1]) might trigger cysteine oxidation. SUB1A-1 may provide enhanced responsiveness to submergence and drought in rice in part due to the fact that it is not a substrate of the N-end rule pathway. By contrast, the condition-dependent destabilization of group VII ERFs in *Arabidopsis* could require oxygen levels to decline below some threshold before these factors can activate anaerobic gene transcription. It is probable that SUB1A-1 evades the N-end rule pathway due to the absence of an optimally positioned lysine downstream of the N degron, as substrate quality is determined combinatorially by an N degron destabilizing residue and downstream lysine position[8,9,25]. Alternatively, differences in protein tertiary structure may preclude

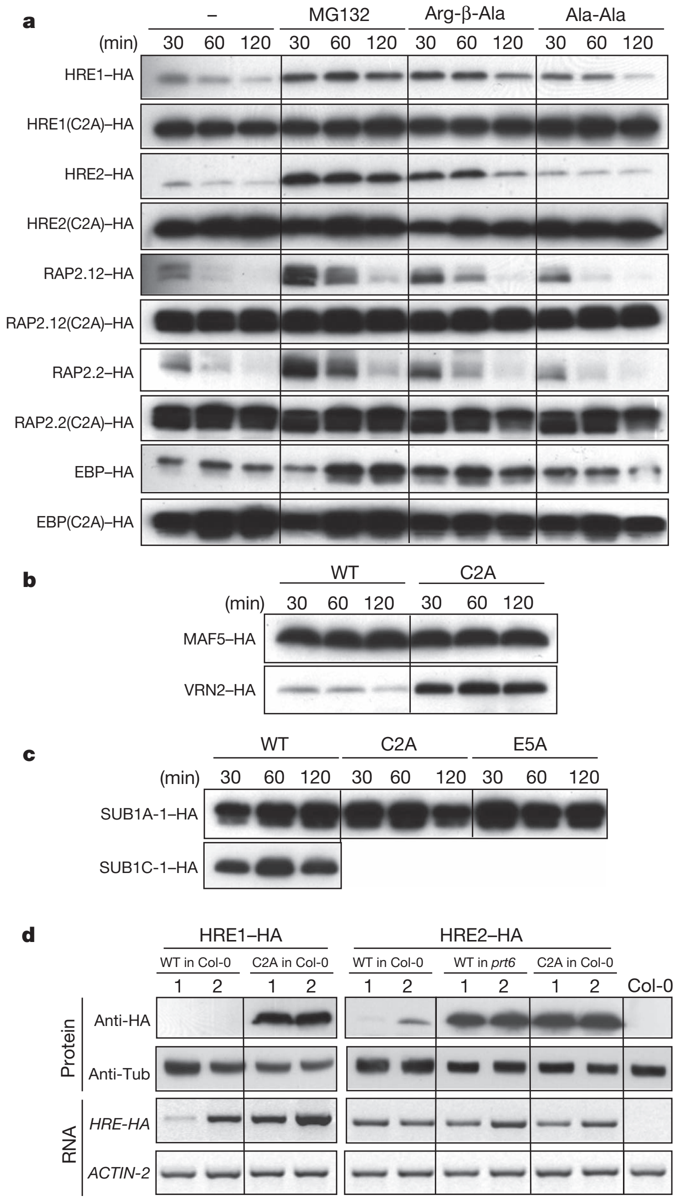


Figure 2 | Group VII ERF transcription factors are substrates for the N-end rule pathway *in vitro* and *in vivo*. **a**, Western blot analysis of *in vitro* stability of HA-tagged wild-type and C2A variants of *Arabidopsis* group VII ERFs in the absence or presence of MG132, N-end rule pathway competitive dipeptide (Arg-β-Ala) or non-competitive dipeptide (Ala-Ala). **b**, *In vitro* stability of wild-type and C2A VRN2–HA and MAF5–HA. **c**, *In vitro* stability of HA-tagged rice ERFs. **d**, *In vivo* protein stability and RNA expression levels of wild-type and C2A variants of HRE1–HA and HRE2–HA ectopically expressed in *Arabidopsis*, shown for two independent transformed lines (1 and 2).

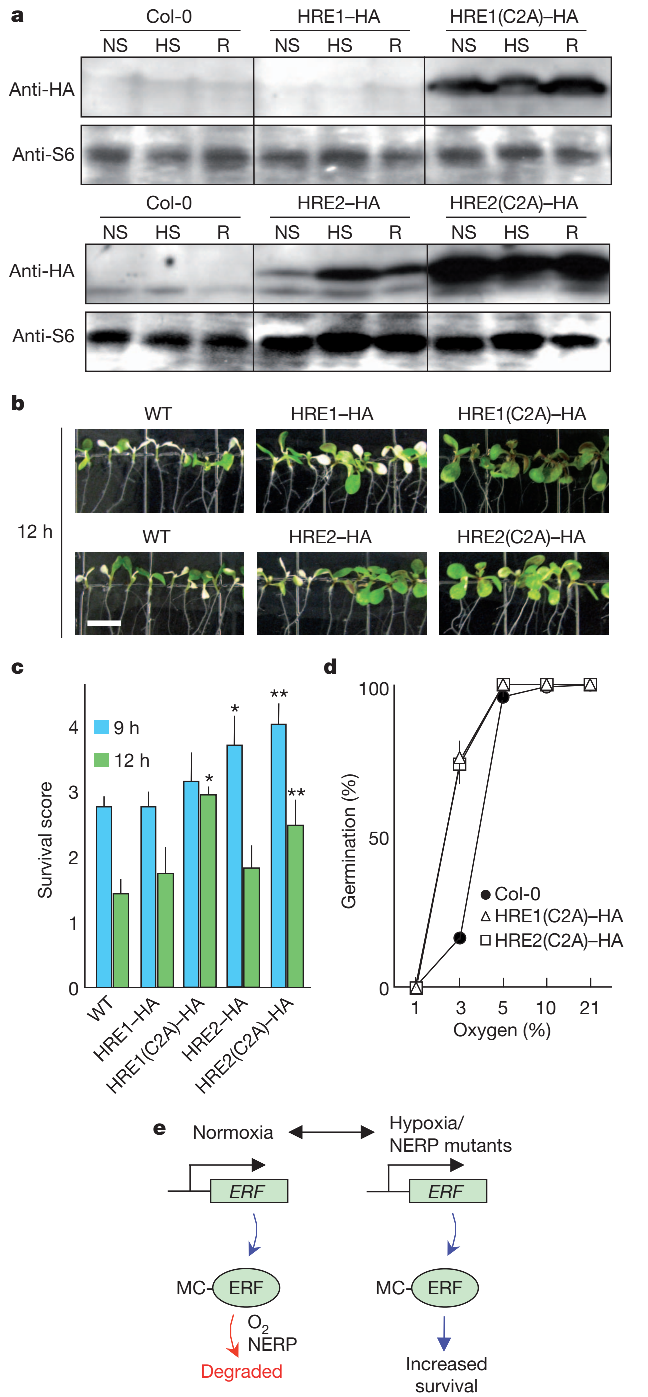


Figure 3 | HRE proteins are stabilized under low oxygen levels and confer hypoxia tolerance. **a**, *In vivo* stability of wild-type and C2A HRE1–HA, HRE2–HA (anti-HA) or S6 (ribosomal protein S6) control (anti-S6). HS, 2 h hypoxia; NS, no stress; R, following 1 h recovery from stress. **b**, Seedlings expressing wild-type or C2A HRE1–HA and HRE2–HA after 12 h hypoxic stress and 3 d of recovery. Scale bar: 0.6 cm. **c**, Seedling survival for wild-type or C2A HRE1–HA and HRE2–HA after 9 h or 12 h hypoxic stress. Data are mean of replicate experiments ± s.d. $*P < 0.05$; $**P < 0.01$. **d**, Germination under reduced oxygen availability. **e**, Model explaining N-end-rule-pathway-mediated oxygen-dependent turnover of group VII ERFs in *Arabidopsis*.

N-terminus accessibility. SUB1A-1 was also recently shown to mediate crosstalk between submergence and drought tolerance in rice by augmenting ABA responsiveness[4], suggesting a link between drought tolerance and the previously identified function of the N-end rule pathway in removing responsiveness to ABA[12]. Targeted degradation of proteins by the N-end rule pathway was identified as a homeostatic mechanism in mammalian systems[22,23,28], for example in the control of hypoxia-related expression of RGS4 (ref. 28) and RGS5 (ref. 23). It is fascinating that the N-end rule pathway carries out the same functionality in relation to low oxygen stress in plants, but taking as substrates members of a plant-specific transcription factor family. This highlights evolutionary conservation of the mechanism of oxygen perception across kingdoms using the N-end rule pathway independent of the targets. Our confirmation of *in vivo* function of two members of ERF group VII provides direct evidence for the control of HRE2 by oxygen and the N-end rule pathway and indirect evidence that HRE1 is also an N-end rule pathway substrate *in vivo*. We demonstrate that all members of *Arabidopsis* group VII ERFs are N-end rule pathway substrates *in vitro*, and thus it is possible that all members orchestrate N-end-rule-pathway-controlled, hypoxia-related functions. Identification and

manipulation of N-end rule pathway substrates will therefore be a key target for both conventional breeding and biotechnological approaches in relation to manipulation of plant responses to abiotic stress.

METHODS SUMMARY

Protein stability analyses. Full-length cDNAs were amplified by polymerase chain reaction (PCR) from either *Arabidopsis thaliana* or *Oryza sativa* L. (cv. M202(*Sub1*)). N-terminal mutations were introduced using the forward primer (Supplementary Table 2). For *in vitro* assays, cDNAs were cloned into a modified version of the pTNT vector (Promega) to produce C-terminal HA fusions. Stability assays were performed using the TNT T7 Coupled Reticulocyte Lysate system (Promega), essentially as described previously[23]. For *in vivo* analysis of HRE–HA proteins, cDNAs were cloned into pE2c, mobilized into pB2GW7 and transformed into *Arabidopsis* using the floral dip method. To assess relative protein stability, equal amounts of total protein extracted from 7-day-old T_3 homozygous seedlings were analysed by western blot, and cDNA synthesized from total RNA was used as a template for semi-quantitative PCR.

Gene expression analyses. For microarray analysis, total RNA extracted from seeds[12] or seedlings[15] was hybridized against the *Arabidopsis* ATH1 genome array (Affymetrix). Differentially expressed genes were clustered as described previously[15]. *pADH::GUS*[16] was crossed to *prt6-1* and homozygous seeds or seedlings were analysed for GUS activity before and after submergence for the times indicated.

Low O_2 phenotypic analyses. To assess germination (scored as radicle emergence), imbibed seeds were incubated for 7-days in chambers flushed with varying O_2 tensions[29]. For 7-day-old seedling survival, O_2 deprivation was achieved by bubbling 99.995% argon through water into chambers under positive pressure, before recovering in air for 3 days and scoring of plants ($n = 15$) per plate that were non-damaged, damaged or dead (scored 5, 3 and 1, respectively)[15]. The same argon chambers were used to treat seedlings for the times indicated before protein extraction for western blot analysis.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 June; accepted 5 September 2011.
Published online 23 October 2011.

1. Bailey-Serres, J. & Voesenek, L. A. C. J. Flooding stress: Acclimations and genetic diversity. *Annu. Rev. Plant Biol.* **59,** 313–339 (2008).
2. Kaelin, W. G. & Ratcliffe, P. J. Oxygen sensing by metazoans: The central role of the HIF hydroxylase pathway. *Mol. Cell* **30,** 393–402 (2008).
3. Xu, K. *et al. Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* **442,** 705–708 (2006).
4. Fukao, T., Yeung, E. & Bailey-Serres, J. The submergence tolerance regulator SUB1A mediates crosstalk between submergence and drought tolerance in rice. *Plant Cell* **23,** 412–427 (2011).
5. Bachmair, A., Finley, D. & Varshavsky, A. *In vivo* half-life of a protein is a function of its amino-terminal residue. *Science* **234,** 179–186 (1986).
6. Varshavsky, A. Regulated protein degradation. *Trends Biochem. Sci.* **30,** 283–286 (2005).
7. Kwon, Y. T. *et al.* An essential role of N-terminal arginylation in cardiovascular development. *Science* **297,** 96–99 (2002).
8. Graciet, E., Mesiti, F. & Wellmer, F. Structure and evolutionary conservation of the plant N-end rule pathway. *Plant J.* **61,** 741–751 (2010).
9. Graciet, E. & Wellmer, F. The plant N-end rule pathway: structure and functions. *Trends Plant Sci.* **15,** 447–453 (2010).
10. Garzón, M. *et al.* PRT6/At5g02310 encodes an *Arabidopsis* ubiquitin ligase of the N-end rule pathway with arginine specificity and is not the CER3 locus. *FEBS Lett.* **581,** 3189–3196 (2007).
11. Tasaki, T. & Kwon, Y. T. The mammalian N-end rule pathway: new insights into its components and physiological roles. *Trends Biochem. Sci.* **32,** 520–528 (2007).
12. Holman, T. J. *et al.* The N-end rule pathway promotes seed germination and establishment through removal of ABA sensitivity in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **106,** 4549–4554 (2009).
13. Yoshida, S., Ito, M., Callis, J., Nishida, I. & Watanabe, A. A delayed leaf senescence mutant is defective in arginyl-tRNA: protein arginyltransferase, a component of the N-end rule pathway in *Arabidopsis*. *Plant J.* **32,** 129–137 (2002).
14. Graciet, E. *et al.* The N-end rule pathway controls multiple functions during *Arabidopsis* shoot and leaf development. *Proc. Natl Acad. Sci. USA* **106,** 13618–13623 (2009).
15. Mustroph, A. *et al.* Profiling translatomes of discrete cell populations resolves altered cellular priorities during hypoxia in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **106,** 18843–18848 (2009).
16. Chung, H. J. & Ferl, R. J. *Arabidopsis* alcohol dehydrogenase expression in both shoots and roots is conditioned by root growth environment. *Plant Physiol.* **121,** 429–436 (1999).
17. Nakano, T., Suzuki, K., Fujimura, T. & Shinshi, H. Genome-wide analysis of the ERF gene family in *Arabidopsis* and rice. *Plant Physiol.* **140,** 411–432 (2006).
18. Licausi, F. *et al.* HRE1 and HRE2, two hypoxia-inducible ethylene response factors, affect anaerobic responses in *Arabidopsis thaliana*. *Plant J.* **62,** 302–315 (2010).
19. Hinz, M. *et al. Arabidopsis* RAP2.2: An ethylene response transcription factor that is important for hypoxia survival. *Plant Physiol.* **153,** 757–772 (2010).
20. Papdi, C. *et al.* Functional identification of *Arabidopsis* stress regulatory genes using the controlled cDNA overexpression system. *Plant Physiol.* **147,** 528–542 (2008).
21. Hattori, Y. *et al.* The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. *Nature* **460,** 1026–1030 (2009).
22. Hu, R. G. *et al.* The N-end rule pathway as a nitric oxide sensor controlling the levels of multiple regulators. *Nature* **437,** 981–986 (2005).
23. Lee, M. J. *et al.* RGS4 and RGS5 are *in vivo* substrates of the N-end rule pathway. *Proc. Natl Acad. Sci. USA* **102,** 15030–15035 (2005).
24. Takahashi, H. *et al.* A simple and high-sensitivity method for analysis of ubiquitination and polyubiquitination based on wheat cell-free protein synthesis. *BMC Plant Biol.* **9,** 39 (2009).
25. Suzuki, T. & Varshavsky, A. Degradation signals in the lysine-asparagine sequence space. *EMBO J.* **18,** 6017–6026 (1999).
26. Leonard, S. E. & Carroll, K. S. Chemical 'omics' approaches for understanding protein cysteine oxidation in biology. *Curr. Opin. Chem. Biol.* **15,** 88–102 (2011).
27. Felle, H. H. pH regulation in anoxic plants. *Ann. Bot.* **96,** 519–532 (2005).
28. Hu, R. G., Wang, H. Q., Xia, Z. X. & Varshavsky, A. The N-end rule pathway is a sensor of heme. *Proc. Natl Acad. Sci. USA* **105,** 76–81 (2008).
29. Côme, D. & Tissaoui, T. Induction d'une dormance embryonnaire secondaire chez le pommier (*Pirus malus* L.) par des atmosphères très appauvries en oxygène. *Compt. Rendus Acad. Sci.* **266,** 477–479 (1968).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements M.J.H., D.J.G., S.G. and C.S.C. were supported by BBSRC grant BB/G010595/1; G.W.B. by a Marie Curie International Incoming Fellowship; N.M.I. by a MARA PhD fellowship from the Malaysian government; S.C.L., T.F. and J.B.-S. by grants NSF IOS-0750811 and NIFA 2008-35100-04528. We thank S. Liddell. Rothamsted Research receives grant-aided support from the BBSRC.

Author Contributions D.J.G., M.J.H., J.B.-S., F.C. and F.L.T. conceived and designed experiments. D.J.G., S.C.L., N.M.I., S.G., C.S.C., G.W.B., T.F. and F.C. performed the experiments. D.J.G., S.C.L., N.M.I., S.G., C.S.C., G.W.B., T.F., F.C. M.J.H., J.B-S. and F.L.T. analysed the data. M.J.H., D.J.G. and J.B.-S. wrote the manuscript.

Author Information The microarray data reported in this paper are deposited in Gene Expression Omnibus under accession number GSE29941 and are also tabulated in Supplementary Information. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.J.H. (michael.holdsworth@nottingham.ac.uk) or J.B.-S. (serres@ucr.edu).

METHODS

Growth and analysis of plant material. *Arabidopsis thaliana* seeds were obtained from NASC, except for transgenics containing *pADH::GUS* (ref. 16) (a gift from R. Ferl). Columbia-0 (Col-0) was the wild type for all analyses. *prt6-1*, *prt6-5* and *ate1 ate2* mutants were described previously[12,14]. For the generation of transgenic *Arabidopsis* and *in vivo* protein assays, plants were grown vertically on half MS media for 7 days at 22 °C in 150 μmol m^{-2} s^{-1} constant light and transferred to soil after 2 weeks if required. For analysis of seedling O_2 deprivation survival and protein analysis, plants were grown vertically on MS medium (0.43% (w/v) MS salts, 1% (w/v) Suc and 0.4% (w/v) phytagel, pH 5.75) at 23 °C with a 16-h-day (50 μmol m^{-2} s^{-1}) and 8-h-night cycle for 7 d. The rice (*Oryza sativa* L.) *SUB1* introgression line cv. M202(*Sub1*) was grown and submerged before cDNA isolation as described previously[3]. All plant experiments were carried out at least three times.

Analysis of oxygen deprivation response in seeds and seedlings. Seven-day-old *Arabidopsis* seedlings were subjected for specified durations to non-stress (NS) or hypoxia stress (HS) treatments, or subjected to hypoxia stress and returned to ambient air (re-oxygenation; R). For seedling survival, 15 Col-0 and 15 mutant seedlings were grown side by side (3 replicates). Treatments commenced at the end of the 16-h light cycle in open (NS) or sealed (HS) chambers. For HS, 99.995% argon gas was bubbled through water and into the chamber while air was expelled by positive pressure[30]. After treatment, the 15 seedlings per genotype per plate were scored as non-damaged, damaged and dead (scored 5, 3 and 1, respectively) compared to wild-type plants grown on the same plate and results analysed using the students *t*-test, as described previously[31], or seedlings were frozen under liquid nitrogen within 3 min of release before protein extraction.

Germination of *Arabidopsis* seeds (3–4 replicates of $n = 60$–100; scored on day 7 as radicle emergence) was performed at 22 °C under constant light in various oxygen tensions achieved through mixing N_2 and air via capillary tubes according to the apparatus described previously[29].

Wild-type plants carrying the *pADH::GUS* transgene[16] were crossed to *prt6-1* plants and homozygous *prt6-1 pADH::GUS* individuals were identified in the F_2 population. Seven-day-old seedlings were submerged in degassed water in the dark to induce hypoxia for the times indicated. Embryos were dissected 6 h after being imbibed. Seedlings and embryos were assayed for GUS activity and imaged following standard methods[32].

Construction of transgenic plants and protein and RNA extractions. To generate C-terminally HA-tagged ERF fusions of *HRE1* (At1g72360) and *HRE2* (At2g47520) driven by the 35SCaMV promoter, full-length cDNAs amplified from *Arabidopsis* total seedling cDNA were first ligated into the Entry vector pE2c and then mobilized into the Destination binary vector pB2GW7, as described previously[33]. N-terminal mutations were incorporated by changing the forward primer sequences accordingly (Supplementary Table 2). Transformation into *Agrobacterium tumefaciens* (strain GV3101 pMP90) and *Arabidopsis thaliana* was performed according to established protocols[34]. Proteins were extracted from 7-day-old homozygous T_3 seedlings as described[35]. Extracts were quantified using the Bio-Rad DC assay and subjected to anti-HA immunoblot analysis. For semi-quantitative RT–PCR, RNA was extracted using an RNEasy plant mini kit (Qiagen) and converted to cDNA using Superscript III Reverse transcriptase (Invitrogen). PCRs were performed with transgene-specific primers (gene-specific forward, HA-tag reverse) and *ACTIN-2* was amplified for use as a loading control (Supplementary Table 2).

***In vitro* analysis of protein stability.** To generate *Arabidopsis* and rice protein–HA fusions driven by the T7 promoter, cDNAs were PCR amplified from *Arabidopsis* total cDNA or submerged rice cDNA (M202(*Sub1*)), as described[3], and ligated into a modified version of the pTNT (Invitrogen) expression vector (pTNT3xHA). N-terminal mutations were incorporated by changing the forward primer sequences accordingly (Supplementary Table 2).

Proteins were expressed *in vitro* using the TNT T7 Coupled Reticulocyte Lysate system (Promega) according to manufacturer's guidelines, using 500 ng plasmid template. Where appropriate, 100 μM MG132 or 1 mM dipeptides (Arg-β-Ala or Ala-Ala; Sigma-Aldrich) and 150 nm Bestatin (Sigma-Aldrich) were added. Reactions were incubated at 30 °C, and samples were taken at indicated time points before mixing with protein loading dye to terminate protein synthesis. Equal amounts of each reaction were subjected to anti-HA immunoblot analysis. All blots were checked for equal loading by Ponceau staining.

Immunoblotting. Proteins resolved by SDS–PAGE were transferred to PVDF using a MiniTrans-Blot electrophoretic transfer cell (Bio-Rad). Membranes were probed with primary antibodies at the following titres: anti-HA (Sigma-Aldrich), 1:1,000; anti-α-tubulin (Sigma-Aldrich), 1:5,000; anti-ribosomal protein S6 (ref. 36), (1:5,000). HRP-conjugated anti-mouse secondary antibody (Santa Cruz) was used at a titre of 1:10,000. Immunoblots were developed to film using ECL western blotting substrate (Pierce).

Alignment of MC-ERF proteins from *Arabidopsis* and rice. Rice and *Arabidopsis* ERF proteins starting with the sequence MC were aligned and phylogenetic relationships observed using CLUSTALW[37].

Microarray hybridization and data analyses. Total RNA extracted from seeds or seedlings was assessed for quality using the Agilent 2100 Bioanalyser with the RNA 6000 Nano reagent kit. Biotin-labelled cRNA was synthesized using the Affymetrix 3′ IVT Express Labelling kit and hybridized against the *Arabidopsis* ATH1 genome array (GeneChip System, Affymetrix). CEL file data were processed to estimate the abundance of each expressed mRNA in two (seedling) or three (imbibed seed) biological replicate samples as described previously[15]. The microarray experiments reported here are described following MIAME guidelines and are deposited in GEO under the accession number GSE29941.

The differentially expressed genes were further analysed by use of fuzzy *k*-means clustering with the FANNY function from the Cluster package in R, as described[15]. The resulting gene-to-cluster assignments are given in Supplementary Table 1 and were visualized with the TIGR MEV program. Each gene cluster was evaluated for enrichment of specific gene functions (Gene Ontology (GO)) as described previously[38] using *Arabidopsis* gene-to-GO mappings from TAIR (http://geneontology.org; downloaded 17 May 2011).

30. Branco-Price, C., Kaiser, K. A., Jang, C. J. H., Larive, C. K. & Bailey-Serres, J. Selective mRNA translation coordinates energetic and metabolic adjustments to cellular oxygen deprivation and reoxygenation in *Arabidopsis thaliana*. *Plant J.* **56,** 743–755 (2008).
31. Mustroph, A. *et al.* Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. *Plant Physiol.* **152,** 1484–1500 (2010).
32. Lucas, M. *et al.* SHORT-ROOT regulates primary, lateral, and adventitious root development in *Arabidopsis*. *Plant Physiol.* **155,** 384–398 (2011).
33. Dubin, M. J., Bowler, C. & Benvenuto, G. A modified Gateway cloning strategy for overexpressing tagged proteins in plants. *Plant Methods* **4,** 3 (2008).
34. Swarup, R. *et al.* Root gravitropism requires lateral root cap and epidermal cells for transport and response to a mobile auxin signal. *Nature Cell Biol.* **7,** 1057–1065 (2005).
35. Martinez-Garcia, J. F., Monte, E. & Quail, P. H. A simple, rapid and quantitative method for preparing *Arabidopsis* protein extracts for immunoblot analysis. *Plant J.* **20,** 251–257 (1999).
36. Williams, A. J., Werner-Fraczek, J., Chang, I. F. & Bailey-Serres, J. Regulated phosphorylation of 40S ribosomal protein S6 in root tips of maize. *Plant Physiol.* **132,** 2086–2097 (2003).
37. Thompson, J. D., Higgins, D. G. & Gibson, T. J. Clustal-W - Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).
38. Horan, K. *et al.* Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol.* **147,** 41–57 (2008).

doi:10.1038/nature10536

Oxygen sensing in plants is mediated by an N-end rule pathway for protein destabilization

Francesco Licausi[1,2], Monika Kosmacz[1], Daan A. Weits[1], Beatrice Giuntoli[2], Federico M. Giorgi[1], Laurentius A. C. J. Voesenek[3,4], Pierdomenico Perata[2] & Joost T. van Dongen[1]

The majority of eukaryotic organisms rely on molecular oxygen for respiratory energy production[1]. When the supply of oxygen is compromised, a variety of acclimation responses are activated to reduce the detrimental effects of energy depletion[2–4]. Various oxygen-sensing mechanisms have been described that are thought to trigger these responses[5–9], but they each seem to be kingdom specific and no sensing mechanism has been identified in plants until now. Here we show that one branch of the ubiquitin-dependent N-end rule pathway for protein degradation, which is active in both mammals and plants[10,11], functions as an oxygen-sensing mechanism in *Arabidopsis thaliana*. We identified a conserved amino-terminal amino acid sequence of the ethylene response factor (ERF)-transcription factor RAP2.12 to be dedicated to an oxygen-dependent sequence of post-translational modifications, which ultimately lead to degradation of RAP2.12 under aerobic conditions. When the oxygen concentration is low—as during flooding—RAP2.12 is released from the plasma membrane and accumulates in the nucleus to activate gene expression for hypoxia acclimation. Our discovery of an oxygen-sensing mechanism opens up new possibilities for improving flooding tolerance in crops.

Tolerance to submergence and low oxygen availability (hypoxia) have been considered to be influenced by different members of subgroup VII of the ERF transcription factor family in *Arabidopsis* (RAP2.12 (ref. 12), RAP2.2 (ref. 13); HRE1 and HRE2 (ref. 14)) and rice (SUB1 (ref. 15), SK1 and SK2 (ref. 16)). Here, we reveal the mechanism by which molecular oxygen acts upon RAP2.12 (At1g53910) to trigger molecular acclimation responses. RAP2.12 is highly homologous to RAP2.2 and is widely conserved in higher plants (Supplementary Fig. 1). It is constitutively expressed throughout the entire plant (Supplementary Fig. 2) and further upregulated in leaves upon hypoxia, but not by the ethylene precursor 1-aminocyclopropane-1-carboxylic-acid (ACC) (Supplementary Fig. 3). RAP2.12 positively regulates gene transcription *in planta* via a conserved carboxy-terminal motif (Supplementary Fig. 4). Constitutive overexpression of RAP2.12 (*35S::RAP2.12*) did not significantly affect the phenotype of *Arabidopsis* plants when grown aerobically (Fig. 1a, b). However, submergence tolerance of independently transformed *35S::RAP2.12* plants increased with respect to the wild-type control, as demonstrated by the increased number and dry weight of plants that recovered from submergence (Fig. 1a, c, d), which can be explained by the faster and stronger induction of hypoxia-responsive genes during the flooding treatment in *35S::RAP2.12* plants (Supplementary Fig. 5). Interestingly, different flooding-tolerance strategies in two wild *Rumex* species correlated with the differential induction of *ERF1*, which is the orthologue of *RAP2.12* (Supplementary Fig. 6). In contrast, constitutive expression of RAP2.12 with a haemagglutinin (HA)-peptide tag at its N terminus (*35S::HA::RAP2.12)* resulted in a reduction of plant growth in air (Fig. 1a, b). Concomitantly, tolerance to submergence decreased as compared to the wild type (Fig. 1c). Similar results were observed when a version of RAP2.12 was expressed from which the first 13 amino acid residues were deleted (*35S::Δ13RAP2.12*). It thus seemed that manipulating the N-terminal amino acid sequence obstructed the regulative function of RAP2.12 already under aerobic conditions, thereby reducing the vigour and stress tolerance of the plants.

To understand the impact of the N-terminal modifications on the activity of RAP2.12, we investigated which genes are expressed under the control of RAP2.12. We found that under aerobic conditions

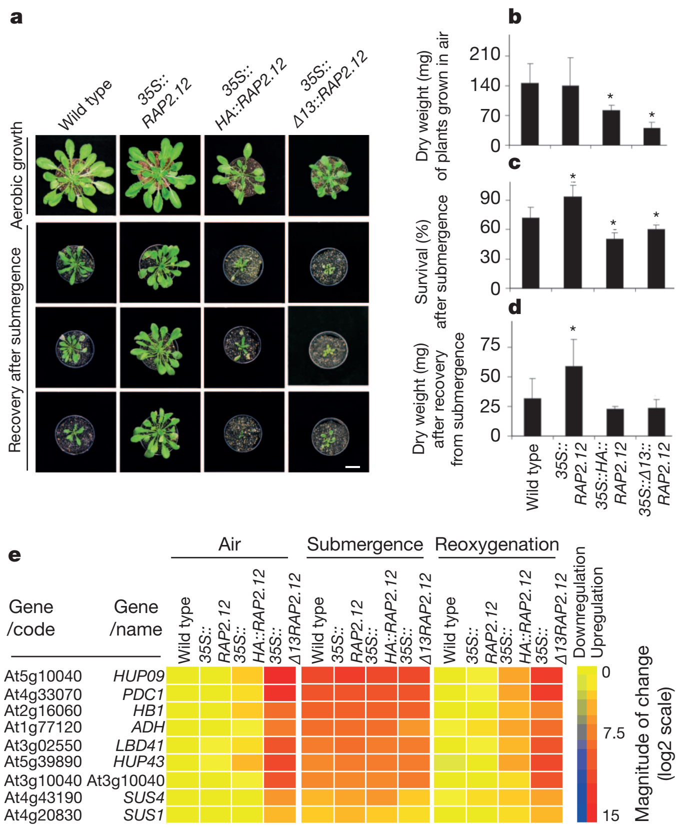


Figure 1 | The transcription factor RAP2.12 regulates hypoxia tolerance of plants. a, The effect of overexpression of RAP2.12, HA::RAP2.12 or Δ13RAP2.12 on plant growth in air, or after submergence. Scale bar, 2 cm. **b**, Dry weight of 7-week-old rosette leaves from air-grown plants ($n = 20$). **c**, Percentage of plants surviving flooding-induced hypoxia ($n = 4$). **d**, Dry weight of rosette leaves from surviving plants, 2 weeks after the flooding treatment ($n = 20$). **e**, Differential expression of hypoxia-responsive genes (reference: wild type in air). Numeric expression values are shown in Supplementary Table 1. Data are presented as mean ± s.d. $^*P < 0.05$, one-way ANOVA.

[1]Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476, Potsdam-Golm, Germany. [2]PlantLab, Institute of Life Sciences, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà 33, 56127 Pisa, Italy. [3]Plant Ecophysiology, Institute of Environmental Biology, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands. [4]Centre for Biosystems Genomics, 6708 PB Wageningen, the Netherlands.

35S::RAP2.12 plants exhibited a slight increase in the expression of hypoxia marker genes, whereas during flooding the expression of these hypoxia marker genes was more strongly upregulated in plants overexpressing RAP2.12 as compared to wild-type plants during flooding (Fig. 1e and Supplementary Table 1). During re-oxygenation, the expression of the hypoxia marker genes was rapidly downregulated in both wild-type plants and *35S::RAP2.12*, whereas in *35S::HA::RAP2.12* and *35S::Δ13RAP2.12* the level of expression remained high, as before the flooding treatment. The correlation between this expression pattern of hypoxia response genes and the reduced growth and recovery after flooding that is observed for plants overexpressing *HA::RAP2.12* and *Δ13RAP2.12* indicates that proper upregulation of the hypoxia response genes during flooding as well as downregulation of these genes during recovery from flooding are both required for optimal plant acclimation. The activation of hypoxic gene expression by RAP2.12 was further confirmed by the observation that RAP2.12 induced a luciferase (*Luc*) reporter gene when its promoter contained the motif ATCTA (Supplementary Fig. 7), which was previously identified as a hypoxia-responsive element in plants[17]. On the other hand, the relatively small effect of *35S::RAP2.12* on gene expression under aerobic conditions (Fig. 1e) indicated that an additional regulatory mechanism reliant on sensing of low oxygen concentrations is needed to induce hypoxic gene expression. Interestingly, this requirement was abolished when the N terminus of RAP2.12 was modified either by fusing the HA-peptide tag to the protein (*35S::HA::RAP2.12* in Fig. 1e), or by deleting its first conserved amino acid residues (*35S::Δ13RAP2.12* in Fig. 1e and Supplementary Fig. 8), indicating that the N terminus of RAP2.12 has an important role in the regulation of the oxygen-dependent activation of the transcription factor.

Further comparative analysis of a full-genome expression profile of the hypoxic response in wild-type plants and the differential regulation of genes by expressing the *HA::RAP2.12* construct under aerobic conditions revealed that the genes that were most strongly up- or downregulated by HA::RAP2.12 were also differentially expressed under hypoxia (Supplementary Fig. 9 and Supplementary Table 2). Similarly, the silencing of *RAP2.12* and its closest homologue *RAP2.2* using an artificial microRNA approach reduced the induction of hypoxic gene expression by low oxygen (Supplementary Fig. 10 and Supplementary Tables 3 and 4). Given that the messenger RNA stability of *RAP2.12* was not affected by the additional nucleotides encoding the N-terminal peptide tag (Supplementary Fig. 11), we concluded that post-translational modifications of the N-terminal amino acid residues of RAP2.12 are involved in regulating the activity of this transcription factor, which is required to induce hypoxia core-response genes.

We further investigated the role of the N-terminal amino acid residues by determining the subcellular localization of RAP2.12 fused to green fluorescent protein (GFP). Under aerobic conditions, the fusion protein localized to the plasma membrane; however, upon hypoxia it accumulated in the nucleus (Fig. 2a and Supplementary Fig. 12). Remarkably, upon re-oxygenation the RAP2.12::GFP signal fully disappeared within 1 h (Fig. 2a). After deleting the conserved N-terminal amino acid residues of RAP2.12 (*35S::Δ13RAP2.12::GFP*), the transcription factor was observed in both the cell membrane and the nucleus under aerobic conditions (Fig. 2a). However, under hypoxia, the

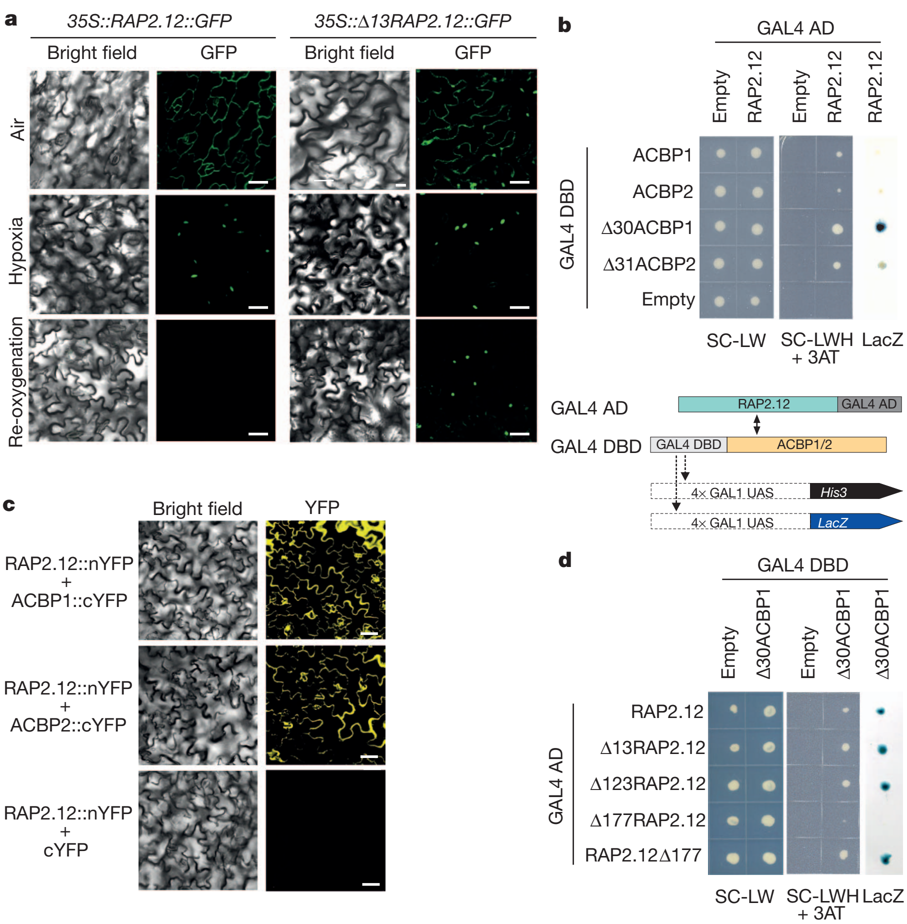


Figure 2 | RAP2.12 is membrane localized and re-localizes in the nucleus upon hypoxia. **a**, Subcellular localization of stably transformed GFP-fused RAP2.12 and Δ13RAP2.12. Localization controls are shown in Supplementary Fig. 12. **b**, Yeast two-hybrid analysis showing interaction between RAP2.12 and ACBP1 and ACBP2. **c**, Bimolecular fluorescence complementation of YFP confirming interaction between RAP2.12 and ACBP1 and ACBP2. **d**, Yeast two-hybrid analysis between various truncated RAP2.12 proteins and Δ30ACBP1. Names of genes are explained in a pictogram shown in Supplementary Fig. 16. AD, activation domain; DBD, DNA-binding domain; UAS, upstream activator sequence. Scale bars, 10 μm.

membrane association disappeared, the protein accumulated only in the nucleus and remained there even after re-oxygenation (*35S::Δ13RAP2.12::GFP* in Fig. 2a). Thus, manipulation of the conserved N terminus of RAP2.12 seems to affect the oxygen-dependent subcellular localization of the transcription factor and, moreover, stabilizes the protein under aerobic conditions.

As RAP2.12 has no hydrophobic domains that could explain its localization at the plasma membrane, we searched for interaction partners of the transcription factor. Yeast two-hybrid analyses (Fig. 2b) and bimolecular fluorescence complementation (BiFC) analysis (Fig. 2c) revealed an interaction between RAP2.12 and the membrane-localized acyl-CoA-binding proteins ACBP1 and ACBP2 (ref. 18), as had been shown previously for ACBP2 and RAP2.3 (ref. 19). The interaction between RAP2.12 and ACBP depended on an amino acid sequence between position 123 and 177, which covers the RAYD motif, a sequence already known to mediate protein–protein interactions[20] (Fig. 2d).

The essential role of the N-terminal residues of RAP2.12 is further supported by the conservation of the first amino acids in almost all members of ERF subfamily VII (Supplementary Fig. 8). The specific sequence of their conserved N terminus qualifies ERF-VII proteins as candidate substrates of the N-end rule pathway[21,22] (Fig. 3a). According to this pathway the terminal Met is removed from the protein by methionine aminopeptidase (MetAP) when the second amino acid of the protein is Cys[23] (Supplementary Fig. 13 and Supplementary Table 5). Terminal Cys is oxidized to cysteine sulphenic acid in an oxygen-dependent manner before arginine transferase (ATE) conjugates an Arg residue to the protein[10,11]. This triggers subsequent ubiquitination by the ligase PROTEOLYSIS 6 (PRT6)[24] and targets the protein to the proteasome for degradation[25], which can occur in both the cytosol and the nucleus[26]. Transient expression of RAP2.12::GFP in *ate1ate2* or *prt6* knockout plants resulted in accumulation of the transcription factor in the nucleus both during aerobic and hypoxic conditions as well as after re-oxygenation (Fig. 3b), similar to what we observed by deleting the N-terminal sequence (Fig. 2a) or after incubation with the proteasome inhibitor MG132 (Supplementary Fig. 14). Western blot analyses showed that the amount of RAP2.12 increased under hypoxia and decreased again after re-oxygenation in the wild type but not in *ate1ate2* or *prt6* (Fig. 3c). The tolerance to submergence of *ate1ate2* and *prt6* rosette plants was reduced (Supplementary Fig. 15), in line with the negative impact of *35S::HA::RAP2.12* and *35S::Δ13RAP2.12* on survival (Fig. 1a, b). Lastly, exchanging the N-terminal Cys with Ala (*35S::MAG-RAP2.12::GFP*) resulted in a GFP signal in the nucleus, similar to what we observed in any of the

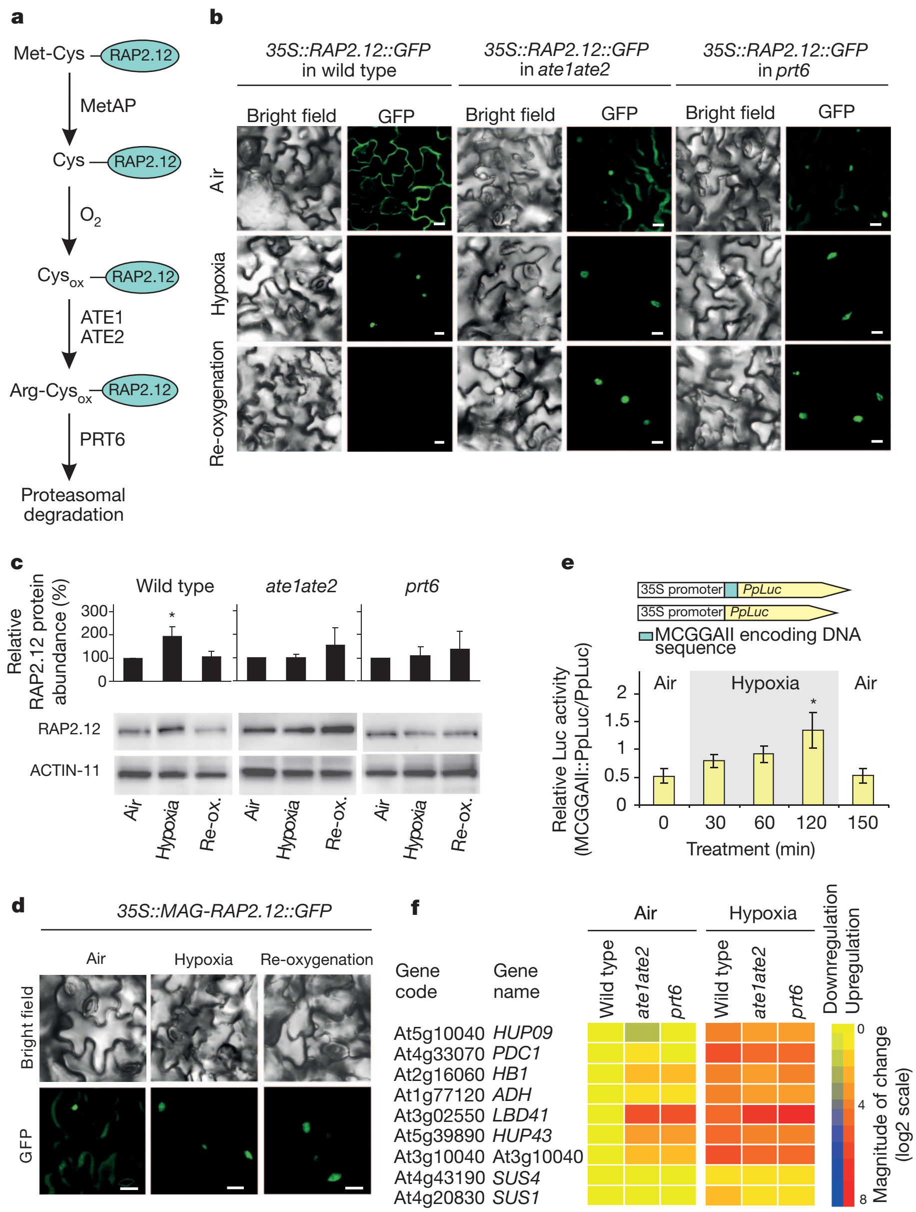


Figure 3 | Oxygen-dependent destabilization of RAP2.12. **a**, Graphical representation of the N-end rule branch that leads to oxygen-dependent protein degradation via the 26S proteasome. **b**, Subcellular localization of transiently expressed RAP2.12::GFP in leaves ($n = 9$). **c**, Western blot and protein quantification of RAP2.12 (means ± s.d., *$P < 0.05$, one-way ANOVA, $n = 3$). Re-ox., re-oxygenation. **d**, Subcellular localization of transiently expressed MAG-RAP2.12::GFP in wild-type leaves ($n = 9$). **e**, Oxygen-dependent modulation of luciferase after fusion with RAP2.12 N-terminal amino acid residues (means ± s.e., *$P < 0.05$, one-way ANOVA, $n = 6$). **f**, Expression of hypoxia-inducible genes in *ate1ate2* and *prt6* mutants.

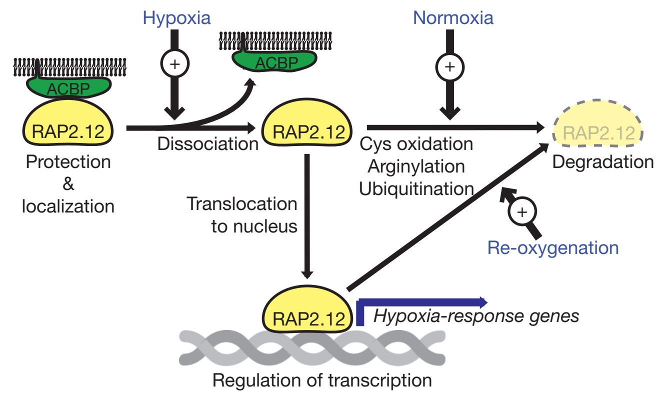


Figure 4 | Model describing the oxygen sensor mechanism in plants. The transcription factor *RAP2.12* is constitutively expressed under aerobic conditions. RAP2.12 protein is always present, bound to ACBP to prevent RAP2.12 from moving into the nucleus under aerobic conditions and to protect it against proteasomal degradation in air. Upon hypoxia, RAP2.12 moves into the nucleus, where it activates anaerobic-gene expression. Upon re-oxygenation, RAP2.12 is rapidly degraded via the N-end rule pathway and proteasome-mediated proteolysis to downregulate the hypoxic response.

other approaches to modify the N-end rule pathway (Fig. 3d). All this indicates that the lifetime of RAP2.12 is controlled by the N-end rule pathway for proteasomal protein degradation.

Next, we investigated whether an oxygen-dependent N-end rule pathway is active in plants and whether it regulates the oxygen-dependent activation of hypoxic gene expression. Fusion of the first conserved N-terminal amino acid residues from RAP2.12 to the Luc reporter protein resulted in an increase of the normalized Luc activity under hypoxic conditions and reduced Luc activity upon re-oxygenation, as predicted by the Cys-oxidation-dependent branch of the N-end rule pathway (Fig. 3e). In addition, constitutive upregulation of hypoxia marker genes was observed under aerobic conditions in plants with reduced ATE and PRT activities (Fig. 3f). This indicates that the oxygen-dependent oxidation of the terminal Cys of RAP2.12 prevents hypoxic gene expression via the destabilization of RAP2.12 in air. Only when the oxygen concentration decreases is Cys oxidation prevented, and the now stably accumulating RAP2.12 can induce the expression of genes involved in the hypoxic response (Fig. 4). Here, we have shown that this oxygen-dependent Cys oxidation is adopted by the ERF-VII factor RAP2.12 and—together with its oxygen-dependent re-localization—triggers the hypoxia-acclimation response in *Arabidopsis*.

METHODS SUMMARY

Unless specifically indicated in the text, low oxygen (hypoxia) conditions used in this study were always maintained at 1% (v/v) oxygen. Full details of the materials and experimental procedures are provided in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 May; accepted 6 September 2011.
Published online 23 October 2011.

1. Webb, J. D., Coleman, M. L. & Pugh, C. W. Hypoxia, hypoxia-inducible factors (HIF), HIF hydroxylases and oxygen sensing. *Cell. Mol. Life Sci.* **66,** 3539–3554 (2009).
2. Kelly, D. P. Hypoxic reprogramming. *Nature Genet.* **40,** 132–134 (2008).
3. Mustroph, A. *et al.* Cross-kingdom comparison of transcriptomic adjustments to low-oxygen stress highlights conserved and plant-specific responses. *Plant Physiol.* **152,** 1484–1500 (2010).
4. Bailey-Serres, J. & Voesenek, L. A. C. J. Flooding stress: acclimations and genetic diversity. *Annu. Rev. Plant Biol.* **59,** 313–339 (2008).
5. Green, J., Crack, J. C., Thomson, A. J. & LeBrun, N. E. Bacterial sensors of oxygen. *Curr. Opin. Microbiol.* **12,** 145–151 (2009).
6. Hou, S. *et al.* Myoglobin-like aerotaxis transducers in Archaea and Bacteria. *Nature* **403,** 540–544 (2000).
7. Osborne, T. F. & Espenshade, P. J. Evolutionary conservation and adaptation in the mechanism that regulates SREBP action: what a long, strange tRIP it's been. *Genes Dev.* **23,** 2578–2591 (2009).
8. Semenza, G. L. HIF-1, O_2, and the 3 PHDs: how animal cells signal hypoxia to the nucleus. *Cell* **107,** 1–3 (2001).
9. van der Wel, H. *et al.* Requirements for Skp1 processing by cytosolic Prolyl 4(*trans*)-hydroxylase and α-*N*-acetylglucosaminyltransferase enzymes involved in O_2 signaling in *Dictyostelium*. *Biochemistry* **50,** 1700–1713 (2011).
10. Lee, M. J. *et al.* RGS4 and RGS5 are *in vivo* substrates of the N-end rule pathway. *Proc. Natl Acad. Sci. USA* **102,** 15030–15035 (2005).
11. Graciet, E., Mesiti, F. & Wellmer, F. Structure and evolutionary conservation of the plant N-end rule pathway. *Plant J.* **61,** 741–751 (2010).
12. Hinz, M. *et al. Arabidopsis* RAP2.2: an ethylene response transcription factor that is important for hypoxia survival. *Plant Physiol.* **153,** 757–772 (2010).
13. Licausi, F. *et al.* HRE1 and HRE2, two hypoxia-inducible ethylene response factors, affect anaerobic responses in *Arabidopsis thaliana*. *Plant J.* **62,** 302–315 (2010).
14. Xu, K. *et al. Sub1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* **442,** 705–708 (2006).
15. Hattori, Y. *et al.* The ethylene response factors *SNORKEL1* and *SNORKEL2* allow rice to adapt to deep water. *Nature* **460,** 1026–1030 (2009).
16. Papdi, C. *et al.* Functional identification of *Arabidopsis* stress regulatory genes using the controlled cDNA overexpression system. *Plant Physiol.* **147,** 528–542 (2008).
17. Licausi, F. *et al.* Hypoxia responsive gene expression is mediated by various subsets of transcription factors and miRNAs that are determined by the actual oxygen availability. *New Phytol.* **190,** 442–456 (2011).
18. Li, H. Y. & Chye, M. L. Membrane localization of *Arabidopsis* acyl-CoA binding protein ACBP2. *Plant Mol. Biol.* **51,** 483–492 (2003).
19. Li, H. Y. & Chye, M. L. *Arabidopsis* Acyl-CoA-binding protein ACBP2 interacts with an ethylene-responsive element-binding protein, AtEBP, via its ankyrin repeats. *Plant Mol. Biol.* **54,** 233–243 (2004).
20. Okamuro, J. K., Caster, B., Villarroel, R., Van Montagu, M. & Jofuku, K. D. The AP2 domain of APETALA2 defines a large new family of DNA binding proteins in *Arabidopsis. Proc. Natl Acad. Sci. USA* **94,** 7076–7081 (1997).
21. Kwon, Y. T. *et al.* An essential role of N-terminal arginylation in cardiovascular development. *Science* **297,** 96–99 (2002).
22. Graciet, E. & Wellmer, F. The plant N-end rule pathway: structure and functions. *Trends Plant Sci.* **15,** 447–453 (2010).
23. Bradshaw, R. A., Brickey, W. W. & Walker, K. W. N-terminal processing: the methionine aminopeptidase and N^{α}-acetyl transferase families. *Trends Biochem. Sci.* **23,** 263–267 (1998).
24. Garzón, M. *et al. PRT6*/At5g02310 encodes an *Arabidopsis* ubiquitin ligase of the N-end rule pathway with arginine specificity and is not the *CER3* locus. *FEBS Lett.* **581,** 3189–3196 (2007).
25. Voges, D., Zwickl, P. & Baumeister, W. The 26S proteasome: a molecular machine designed for controlled proteolysis. *Annu. Rev. Biochem.* **68,** 1015–1068 (2000).
26. Vallon, U. & Kull, U. Localization of proteasomes in plant cells. *Protoplasma* **182,** 15–18 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank H. van Veen and R. Sasidharan (for providing *Rumex* data), E. Maximova, F. Kragler (microscopy), W. Schulze and R. Bock (support and discussion), A. Fernie and R. Pierik (commenting on the manuscript) and S. Parlanti, L. Bartezko and K. Seehaus (plant cultivation). This work was financially supported by the Max Planck Institute of Molecular Plant Physiology, Scuola Superiore Sant'Anna, and the Deutsche Forschungsgemeinschaft (DFG) (DO 1298/2-1).

Author Contributions F.L., M.K., D.A.W. and B.G. performed the experiments. F.M.G. carried out the bioinformatical analysis. F.L., L.A.C.J.V., P.P. and J.T.v.D. designed the experiments. F.L., P.P. and J.T.v.D. wrote the manuscript. All the authors discussed and commented on the content of the paper.

Author Information The raw data files of the microarray experiments have been deposited in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo; accession number: GSE29187). The gene sequences for the *Rumex spp.* used in this work have been deposited at NCBI (*RaERF1*: JF968115; *RaERF2*: JF968116; *RpERF1*: JF968117; *RpERF2*: JF968118; and *RpERF3*: JF968119). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to F.L. (f.licausi@sssup.it) or J.T.v.D. (dongen@mpimp-golm.mpg.de).

METHODS

Plant materials. *A. thaliana* Columbia-0 (Col-0) was used as wild-type ecotype, as described in the figure legends. Double *ate1ate2* knockout seeds were provided by E. Graciet, *prt6* knockout seeds (line EOL4) were obtained from the Institute of Agronomic Research, *35S::GFP* seeds were provided by M. Kawai-Yamada.

Growth conditions and phenotypic evaluation. Seeds were sown in moist soil, stratified at 4 °C in the dark for 48 h and germinated at 22 °C day/18 °C night with a photoperiod of 8 h light and 16 h darkness. For all experiments 5-week-old plants were used. Low oxygen (1% (v/v) oxygen in air) treatments were performed as described previously[17]. Flooding tolerance was assayed using three independent transgenic lines. Plants were submerged with deionized water in 15-cm-high plastic boxes and kept in the dark. Leaves were at 5 cm under the water surface. After 84 h, the water was removed from the boxes and photoperiodic conditions (8 h/16 h, light/dark) were restored. Tolerance assays were repeated four times by using 10–20 plants per genotype each time. *Rumex spp.* cultivation and submergence treatment were performed as described previously[27].

Cloning of the various constructs. Coding sequences (CDSs) were amplified from a cDNA template using Phusion High Fidelity DNA-polymerase (New England Biolabs). An artificial microRNA (amiRNA) against RAP2.12 was generated by overlapping PCR using the pRS300 vector as backbone. All open reading frames were cloned into pENTR/D-TOPO (Invitrogen). The resulting entry vectors were recombined into destination vectors using the LR reaction mix II (Invitrogen) to obtain the expression vectors. A complete list of all destination vectors and primers used is provided in Supplementary Tables 6 and 7, respectively.

Plant transformation. Stable transgenic plants were obtained using the floral dip method[28]. T0 seeds were screened for kanamycin or phosphinotricine resistance and single-insertion lines were identified as described previously[13] Transient leaf transformations using 3-week-old plants were performed as described previously[29]. All transient expression essays were repeated at least three times using independently grown plants. Each time the experiment was repeated, we transformed leaves from three independent plants. So, at least nine independent transformations from at least three different plant cultures were analysed.

qRT–PCR. RNA extraction, removal of genomic DNA, cDNA synthesis and qRT–PCR analyses were performed as described previously[13]. For *35S::RAP2.12*, *35S::HA::RAP2.12*, *amiRAP2.2-12* and *35S::Δ13RAP2.12* three independent transgenic lines were used and the average expression value was calculated. For all the other genotypes, three independent biological replicates were used.

Microarrays. Three independent *RAP2.12* overexpressors or *RAP2.2-RAP2.12* silenced lines were grown in soil for 5 weeks and then subjected to a treatment with 1% oxygen in the dark for 90 min. Total RNA from whole rosettes was extracted as described for the qRT–PCR analyses. Hybridization and scanning procedures were performed by NASC (http://arabidopsis.info/). Microarray analysis and data quality control were performed as described previously[13] using Robin[30]. Normalization of the raw data and an estimation of signal intensities were carried out using the Genechip Robust Multiarray Average (GC-RMA) methodology[31]. Differential gene expression analysis was carried out using limma[32], with a Benjamini–Hochberg *P*-value correction[33]. Microarray data sets were deposited in a public repository with open access (accession number GSE29187; http://www.ncbi.nlm.nih.gov/geo/).

Confocal imaging. For GFP and YFP imaging, leaves from independent stable or transiently transformed 4-week-old plants were analysed with a Leica DM6000B/SP5 confocal microscope (Leica Microsystems).

Reporter transactivation assay. *Arabidopsis* mesophyll protoplasts were used to identify the region responsible for the trans-activation activity of RAP2.12. The DNA-binding domain from *Saccharomyces cerevisiae* was fused at the N terminus of RAP2.12 and its deletion variants. The UAS fused to a minimal 35S promoter was inserted into pGreenII-800LUC to generate a reporter and normalization vector. Non-recombined pDBD-GW vector was used as a negative control. Protoplasts were prepared according to a previously described method[34] and transfected using 5 μg plasmid DNA each. A dual luciferase reporter assay was performed as described previously[16].

Protein stability assay using the Luc reporter system. Leaves of *A. thaliana* Col-0 were transformed with either a *35S::PpLuc* or a *35S::MCGGAII::PpLuc* constructs (both containing also a *35S::RrLuc* cassette for normalization purposes). Normalized luciferase activity (PpLuc/RrLuc) was measured as described previously[17] and Luc protein stability was evaluated as the ratio between *MCGGAII::PpLuc* and *PpLuc* transfected leaves. The experiment was repeated three times using five independent replicates in each repetition.

Yeast two-hybrid assay. The ProQuestTm Two-hybrid System (Invitrogen) was used. PExpTM32/Krev1 and pEXPTM22/RalGDS-wt were used as positive controls, and pDESTTM32 and pDESTTM22 as negative controls. *S. cerevisiae* strain Mav203 was transformed with the different combinations of bait, prey and control vectors (Supplementary Fig. 16). Colonies containing both vectors were selected by plating at 28 °C to select colonies containing an interacting protein partner for 3 days on minimal selective dropout medium lacking Leu and Trp (SC-LW medium). They were subsequently replicated on selective dropout medium (SC-LWH+3AT medium) lacking Leu, Trp, His and supplemented with 10 mM 3-aminotriazole (3AT). The strength of the interaction was further verified by β-galactosidase staining (LacZ) following the manufacturer's instructions.

SC-LW, control medium without Leu and Trp; SC-LWH+3AT, selective medium without Leu, Trp, His and with 3AT.

BiFC. *In planta* protein interactions were investigated with bimolecular fluorescence complementation in an *Arabidopsis* transient expression system as described previously[35].

SDS–PAGE and western blotting. Protein samples from total tissue extracts were separated by SDS–PAGE on 10% acrylamide midigels (Biorad) and then transferred onto a polyvinylidene difluoride membrane (BioRad). Incubations with the antiserum and the secondary antibody conjugated to horseradish peroxidase (Agrisera) were performed following the method recommended for the ECL Plus western blotting detection system (GE Healthcare).

Polyclonal anti-RAP2.12 antibodies were affinity purified at Genscript laboratories after being raised in rabbits against a RAP2.12/RAP2.2 specific synthetic peptide (NLKGSKKSSKNRSN). Lyophilized antibody was re-suspended to an approximate concentration of 1 μg ml^{-1}. A monoclonal antibody against *Arabidopsis* ACTIN-11 (Agrisera, AS10 702) was used to confirm equal loading and transfer. Densitometric analysis of the protein signals on the western blots was performed with the software package UVP VisionWorks LS (Ultra-Violet Products). Normalization was carried out using the ACTIN-11 signal and setting to 100 the relative protein signal value for each of the 'air' controls.

Statistical analyses. Significant variations between genotypes or treatments were evaluated statistically by Sigmaplot using either a *t*-test, one-way or two-way ANOVA where appropriate. Mean values that were significantly different ($P < 0.05$) from the control or wild-type treatment are marked with an asterisk. The statistical evaluation of the microarray experiments is described earlier.

27. Pierik, R., de Wit, M. & Voesenek, L. A. C. J. Growth-mediated stress escape: convergence of signal transduction pathways activated upon exposure to two different environmental stresses. *New Phytol.* **189,** 122–134 (2011).
28. Zhang, X., Henriques, R., Lin, S. S., Niu, Q. W. & Chua, N. H. Agrobacterium-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nature Protocols* **1,** 641–646 (2006).
29. Lee, M. W. & Yang, Y. Transient expression assay by agroinfiltration of leaves. *Methods Mol. Biol.* **323,** 225–229 (2006).
30. Lohse, M. *et al.* Robin: An intuitive wizard application for R-based expression microarray quality assessment and analysis. *Plant Physiol.* **153,** 642–651 (2010).
31. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3,** Article 3 (2004).
32. Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19,** 368–375 (2003).
33. Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99,** 909–917 (2004).
34. Yoo, S. D., Cho, Y. H. & Sheen, J. *Arabidopsis* mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nature Protocols* **2,** 1565–1572 (2007).
35. Gehl, C., Waadt, R., Kudla, J., Mendel, R.-R. & Hänsch, R. New GATEWAY vectors for high throughput analyses of protein–protein interactions by bimolecular fluorescence complementation. *Mol. Plant* **2,** 1051–1058 (2009).

doi:10.1038/nature10537

Structural basis of RNA recognition and activation by innate immune receptor RIG-I

Fuguo Jiang[1]*, Anand Ramanathan[2]*, Matthew T. Miller[1], Guo-Qing Tang[2], Michael Gale Jr[3], Smita S. Patel[2] & Joseph Marcotrigiano[1]

Retinoic-acid-inducible gene-I (RIG-I; also known as DDX58) is a cytoplasmic pathogen recognition receptor that recognizes pathogen-associated molecular pattern (PAMP) motifs to differentiate between viral and cellular RNAs. RIG-I is activated by blunt-ended double-stranded (ds)RNA with or without a 5′-triphosphate (ppp), by single-stranded RNA marked by a 5′-ppp[1] and by polyuridine sequences[2,3]. Upon binding to such PAMP motifs, RIG-I initiates a signalling cascade that induces innate immune defences and inflammatory cytokines to establish an antiviral state. The RIG-I pathway is highly regulated and aberrant signalling leads to apoptosis, altered cell differentiation, inflammation, autoimmune diseases and cancer[4,5]. The helicase and repressor domains (RD) of RIG-I recognize dsRNA and 5′-ppp RNA to activate the two amino-terminal caspase recruitment domains (CARDs) for signalling. Here, to understand the synergy between the helicase and the RD for RNA binding, and the contribution of ATP hydrolysis to RIG-I activation, we determined the structure of human RIG-I helicase-RD in complex with dsRNA and an ATP analogue. The helicase-RD organizes into a ring around dsRNA, capping one end, while contacting both strands using previously uncharacterized motifs to recognize dsRNA. Small-angle X-ray scattering, limited proteolysis and differential scanning fluorimetry indicate that RIG-I is in an extended and flexible conformation that compacts upon binding RNA. These results provide a detailed view of the role of helicase in dsRNA recognition, the synergy between the RD and the helicase for RNA binding and the organization of full-length RIG-I bound to dsRNA, and provide evidence of a conformational change upon RNA binding. The RIG-I helicase-RD structure is consistent with dsRNA translocation without unwinding and cooperative binding to RNA. The structure yields unprecedented insight into innate immunity and has a broader impact on other areas of biology, including RNA interference and DNA repair, which utilize homologous helicase domains within DICER and FANCM.

To investigate the contributions of the individual domains of RIG-I to RNA binding, we used fluorescence anisotropy to determine the equilibrium dissociation constants (K_d) of the protein–RNA complexes. The tightest RNA affinity was observed with the helicase-RD, whereas the full-length RIG-I, helicase domain and RD bind dsRNA with a 24-fold, 8,600-fold and 50-fold weaker affinity, respectively, and all proteins bind RNA with a 1:1 stoichiometry (Table 1 and Supplementary Fig. 1). Consistent with the affinities, full-length RIG-I and helicase-RD demonstrated robust dsRNA-stimulated ATPase activity, whereas the helicase domain showed a weak ATPase activity (Table 1). Interestingly, the presence of a 5′-ppp in dsRNA does not alter the stoichiometry of RNA binding, although the 5′-ppp dsRNA binds more tightly than the 5′-OH dsRNA to full-length RIG-I.

Crystals of RIG-I helicase-RD in complex with ADP•BeF_3 and 14 base-pair palindromic dsRNA diffracted to 2.9 Å resolution (Supplementary Table 1). The dsRNA maintains an A-form helical conformation and interacts with all four domains of helicase-RD, which are arranged into a ring around the dsRNA (Fig. 1a–c). The RecA-like helicase domain (domain 1) progresses to an α-helical domain (domain 3) and into the second RecA-like helicase domain (domain 2) that ends with the RD. The notable features of the RIG-I structure are the linkers that connect the domains. Domains 1 and 3 are connected by a β-strand that is part of the parallel β-sheet in domain 2. The RD is connected to domain 2 via a prominent V-shaped linker, consisting of two α-helices that interact extensively with domains 1 and 2 with a buried surface over 1,500 $Å^2$. The V-shaped linker extends into a proline-rich (^{796}KPKPVPD) loop that makes the final connection to the RD. Interestingly, T770 at the vertex of the V-shaped linker is one of the phosphorylation sites that regulates RIG-I signalling[6].

Molecular surface analysis shows that one end of the dsRNA is capped by the entire helicase-RD molecule whereas the opposite end of the RNA is exposed (Fig. 1d–f). At the capped end, the 5′ terminus of the dsRNA abuts the RD domain but the 3′ terminus is somewhat

Table 1 | Summary of ATPase rates and affinity measurements of dsRNA binding

	ATPase		Affinity measurements*			
	RNA	ATPase activity $M/(M \times s)$	RNA	Nucleotide analogue	Dissociation constant (K_d, nM)	Final anisotropy
RD			14-bp dsRNA-F	No nucleotide	2.6 ± 1.20	0.109
Helicase	14-bp dsRNA	0.89 ± 0.04	14-bp dsRNA-F	No nucleotide	436 ± 121	0.138
	No RNA	Undetectable		ADP•BeF_3	1300 ± 300†	0.138
Helicase-RD	14-bp dsRNA	21.8 ± 0.70	14-bp dsRNA-F	No nucleotide	0.05 ± 0.02	0.213
	14-bp palindromic dsRNA	24.4 ± 0.21	5′-ppp 14-bp dsRNA-F	ADP•BeF_3	0.02 ± 0.07	0.280
	No RNA	Undetectable		No nucleotide	0.03 ± 0.03	0.190
Full-length RIG-I	14-bp dsRNA	28.2 ± 1.42	14-bp dsRNA-F	No nucleotide	1.2 ± 0.6†	0.187
	No RNA	Undetectable	5′-ppp 14-bp dsRNA	No nucleotide	0.44 ± 0.27	0.183

Mean ATPase rate constant from two independent measurements and range is shown. Affinity measurements and range from two or more independent experiments are shown. Initial anisotropy of the free fluorescein-labelled dsRNA was 0.10 for all experiments.
* Data was fitted to a 1:1 (protein:RNA) binding model
† Value from single measurement and fitting error is shown.

[1]Center for Advanced Biotechnology and Medicine, Department of Chemistry and Chemical Biology, Rutgers University, 679 Hoes Lane West, Piscataway, New Jersey 08854, USA. [2]Department of Biochemistry, UMDNJ-RWJ Medical School, 675 Hoes Lane West, Piscataway, New Jersey 08854, USA. [3]Department of Immunology, University of Washington School of Medicine, 1959 NE Pacific Street, Seattle, Washington 98195, USA.
*These authors contributed equally to this work.

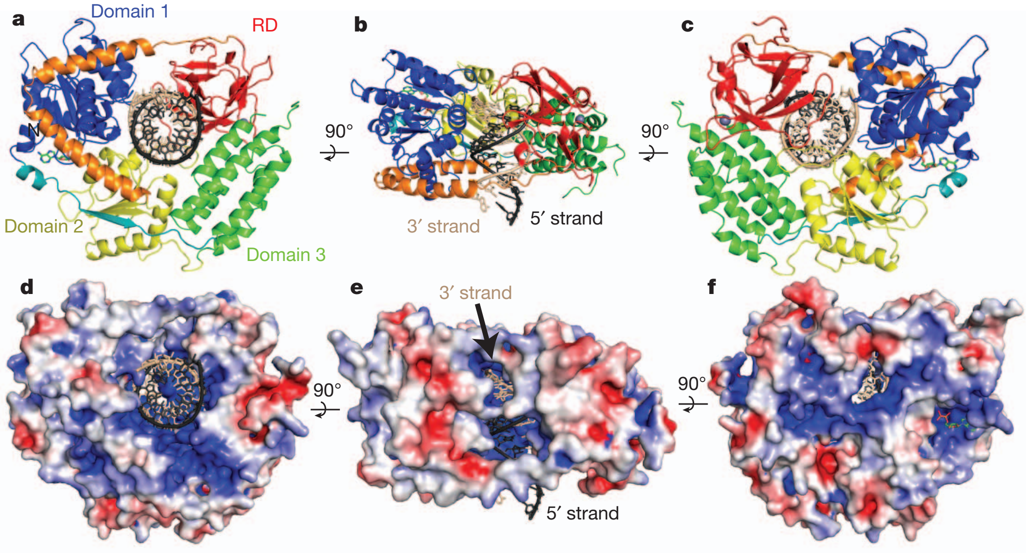


Figure 1 | Structural overview of RIG-I helicase-RD. **a–c**, Schematic representation of the RIG-I helicase-RD, highlighting the RecA-like domain 1 (blue), the α-helical domain 3 (green), the RecA-like helicase domain 2 (yellow) and the RD (red). The linker connecting domain 1 with domain 3 is coloured teal, and the V-shaped linker between domain 2 and the RD is coloured orange. The ADP•BeF_3 and dsRNA are shown in stick representation with the 5′ and 3′ strands of the RNA coloured black and beige, respectively. A grey sphere denotes the position of the zinc ion in the RD. **d–f**, The surface of the RIG-I helicase-RD coloured for electrostatic potential at $\pm 5\,kT\,e^{-1}$: blue (basic), white (neutral) and red (acidic). The views in panels **a**, **b** and **c** are identical to those in **d**, **e** and **f**, respectively.

exposed through a highly basic channel (Fig. 1b, e). The 3′-terminus RNA strand will be referred to as the 3′ strand, and the opposite strand will be referred to as the 5′ strand. This basic channel may allow RIG-I to recognize dsRNA with 3′ nucleotide overhangs or 3′ monophosphates that are products of RNase L digestion[7]. Additionally there are visible channels along the long axis of the dsRNA, which may allow the

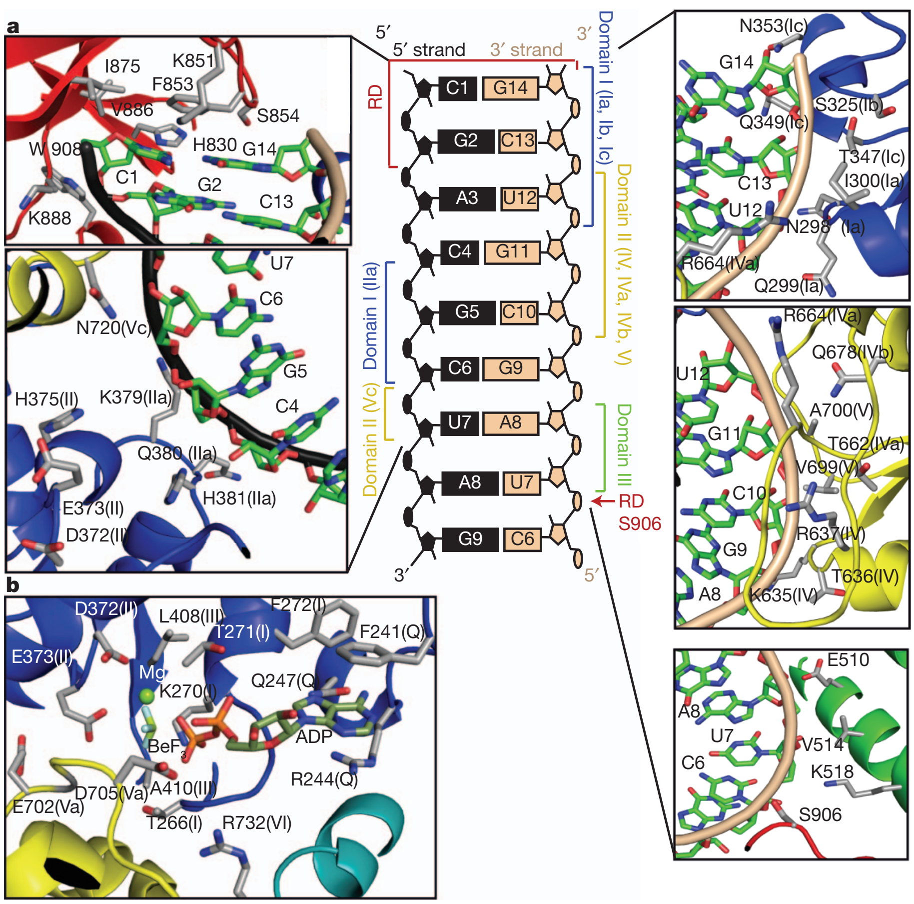


Figure 2 | Interactions of the RIG-I helicase-RD with dsRNA and ADP•BeF_3. **a**, A schematic representation showing the interactions between RIG-I domains and helicase motifs (given in parentheses) with dsRNA located in the centre. Detailed contacts are shown in the surrounding panels. **b**, Stick representation detailing the interactions of RIG-I helicase motifs with ADP•BeF_3 and Mg^{2+} is shown in the lower left panel.

domains to flex and accommodate bulges or noncanonical base pairs (Fig. 1e, f).

All four domains of the helicase-RD participate in dsRNA binding, burying a total of 1,500 Å^2 of surface area to encircle about eight base pairs (Fig. 2a). Most contacts are with the sugar phosphate backbone of both strands but there are a few base-specific interactions. The F853 in the RD stacks over the terminal C1–G14 base pair, and H830 and S854 form hydrogen bonds with the ribose 2′-OH of C1 (5′ strand) and G14 (3′ strand), respectively. Phosphorylation of S854 and S855, shown to negatively regulate RIG-I signalling[6], would be predicted to adversely affect RNA binding. The RD–RNA contacts in the helicase-RD structure are identical to those identified in the structures of the isolated RD bound to dsRNA[8–10]. Residues H847, K858 and K861 of the RD, reported to interact with the 5′-ppp, do not make any new interactions in the helicase-RD structure. The only RD contact with the 3′ strand, other than at the blunt end, is between S906 and the U7 phosphate backbone.

The core helicase (domains 1 and 2) contains characteristic motifs critical for RNA binding and ATP hydrolysis[11] (Supplementary Fig. 3). The two helicase domains together contact five nucleotides from the 3′ end of the 3′ strand (G14–C10) and four nucleotides in the middle of the 5′ strand (C4–U7) (Fig. 2a). Motif Ib (S325 and G326) and motif Ic (T347, Q349 and N353) of domain 1 contact the ribose-phosphate backbone of the terminal base of the 3′ strand (G14), while motif Ia (N298, Q299 and I300) interacts with C13 and U12. Motifs IV, IVa, IVb and V in domain 2 interact with the 3′ strand from C13 to C10. Residues in motif IV (K635, T636 and R637) interact with the phosphate backbone of G11–C10. Similarly, motif IVa (T662, G663 and R664) and motif IVb (Q678) contact the backbone of C13–G11, and motif V (V699) interacts with the RNA backbone at C10.

All RIG-I-like helicases contain a large insertion (domain 3) between the core helicase domains[12]. The structure of the RIG-I-like helicase Hef showed a similar α-helical domain 3, except it is rotated in RIG-I owing to dsRNA interactions (Supplementary Fig. 4). An α-helix within domain 3 (residues 506–522) of RIG-I runs almost perpendicular to the minor groove of the dsRNA and interacts with the 3′ strand without contacting the 5′ strand (Fig. 2a). The interactions of domain 3 with the dsRNA via residues E510, V514 and K518 extend the helicase contacts to U7 and A8 of the 3′ strand.

Although most interactions of the RIG-I are with the 3′ strand, all three helicase domains are in close proximity to the 5′ strand and two new motifs contact the 5′ strand. Motif IIa (^{379}KQHPY) immediately follows motif II (^{372}DECH) and interacts with C4–C6 of the 5′ strand. Similarly, N720 in motif Vc in domain 2 contacts the 5′-strand backbone at C6 and U7. These new motifs may represent a general feature of helicases binding to dsRNA or dsDNA, as a region similar to motif IIa was identified in the Swi/Snf2 family helicases and a Rad54 homologue dsDNA complex structure[13].

RIG-I helicase-RD contains one ADP•BeF_3 molecule bound at the interface of domains 1 and 2 via the conserved helicase motifs Q, I, II, III, Va and VI, which are generally involved in ATP binding/hydrolysis (Fig. 2b)[12]. The Q motif makes adenine-specific contacts, whereas the other motifs are involved in binding the triphosphate moiety and the Mg^{2+}. Motif I (^{267}GCGKT) contacts the phosphates and the BeF_3, which mimics the γ-phosphate of ATP as observed in other SF2 helicases[12]. The D372 and E373 of motif II (^{372}DECH) coordinate the Mg^{2+} to stabilize the ATP analogue. The helicase motifs Va and VI, which contact the ribose and phosphates of the ATP in other helicases, are close to ADP•BeF_3.

RIG-I has been shown to have translocation activity on dsRNA[14] and has been reported to unwind short dsRNA[15], although we (data not shown) and others[14] have failed to detect unwinding activity. Superposition of RIG-I with hepatitis C virus (HCV) NS3 helicase (NS3h) bound to single-stranded (ss)DNA[16] (Fig. 3a) makes predictions about translocation and lack of helicase activity of RIG-I. First, ssDNA in NS3h overlays with the 3′ strand of the dsRNA bound to

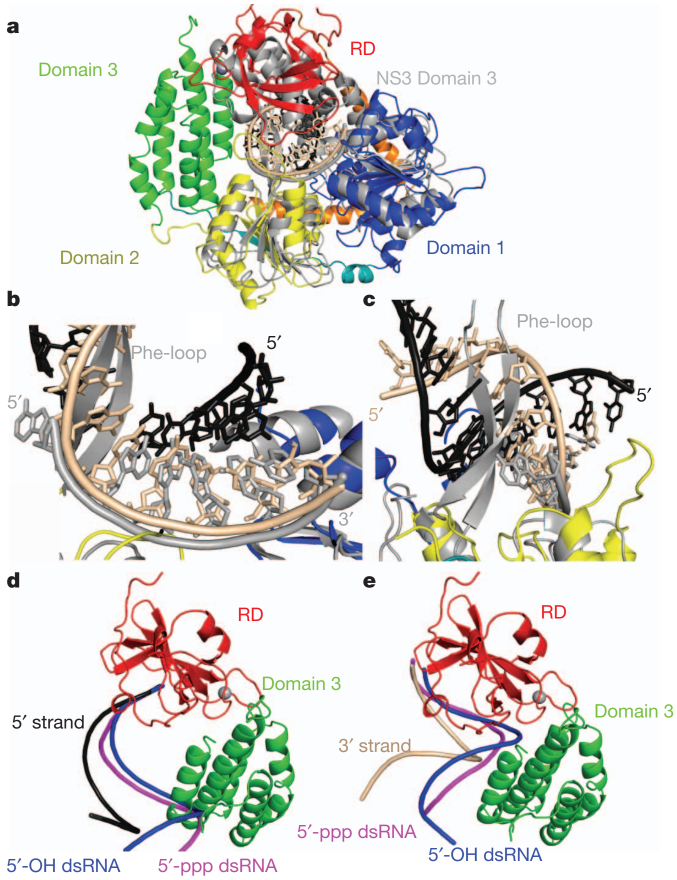


Figure 3 | Comparison of RIG-I helicase-RD with HCV NS3h and RD bound to 5′-OH and 5′-ppp dsRNA. **a**, Ribbon diagram showing the superposition of the RIG-I helicase-RD–dsRNA–ADP•BeF_3 structure and NS3h bound to ssDNA (PDB accession 3KQH) (grey). The helicase core domains 1 and 2 from RIG-I helicase-RD superimpose well, whereas domain 3 of NS3h is positioned over the RD. **b**, Superposition of the RIG-I helicase-RD with NS3h demonstrates that the ssDNA bound to NS3h overlays with the 3′ strand (beige) of the dsRNA bound to the helicase-RD. **c**, The location of the Phe-loop of NS3h relative to the dsRNA of the RIG-I helicase-RD–dsRNA–ADP•BeF_3 structure. **d**, **e**, Superposition of the 5′-OH (blue; PDB accession 3OG8) and 5′-ppp dsRNA (magenta; PDB accession 3LRR) based on the location of the RD. For clarity the 5′ strands (**d**) and 3′ strands (**e**) are shown separately.

RIG-I, suggesting that the RIG-I helicase makes principal motor contacts with the 3′ strand (Fig. 3b) when translocating along dsRNA. Second, a conserved Phe-loop in NS3h bisects the dsRNA, consistent with its implicated role in unwinding[17] (Fig. 3b, c). The absence of such a motif in RIG-I could explain the lack of RNA-unwinding activity. Interestingly, the RIG-I RD positions over domain 3 of NS3h, indicating that this domain of NS3h may interact with the 5′ strand of dsRNA (Fig. 3a).

The RD superimposes well with a root mean squared deviation of ~0.6 Å with the structures of isolated RD bound to dsRNA with and without 5′-ppp[8–10]. However, the trajectory of the RNA helix in the helicase-RD is different. Superposition of the three RD structures demonstrates that the dsRNA bound to the isolated RD clashes with domain 3 (Fig. 3d, e). Thus, the dsRNA in helicase-RD is rotated and the previously reported contacts (R811, K849, K851 and H871) between RD and RNA are no longer observed in the helicase-RD structure.

It has been proposed that RIG-I exists in an auto-inhibited state[18] with undetectable ATPase activity (Table 1) in the absence of RNA, and activates upon binding viral RNAs. Small-angle X-ray scattering (SAXS), limited proteolysis and differential scanning fluorimetry (DSF) were performed to gain insights into how RNA binding activates

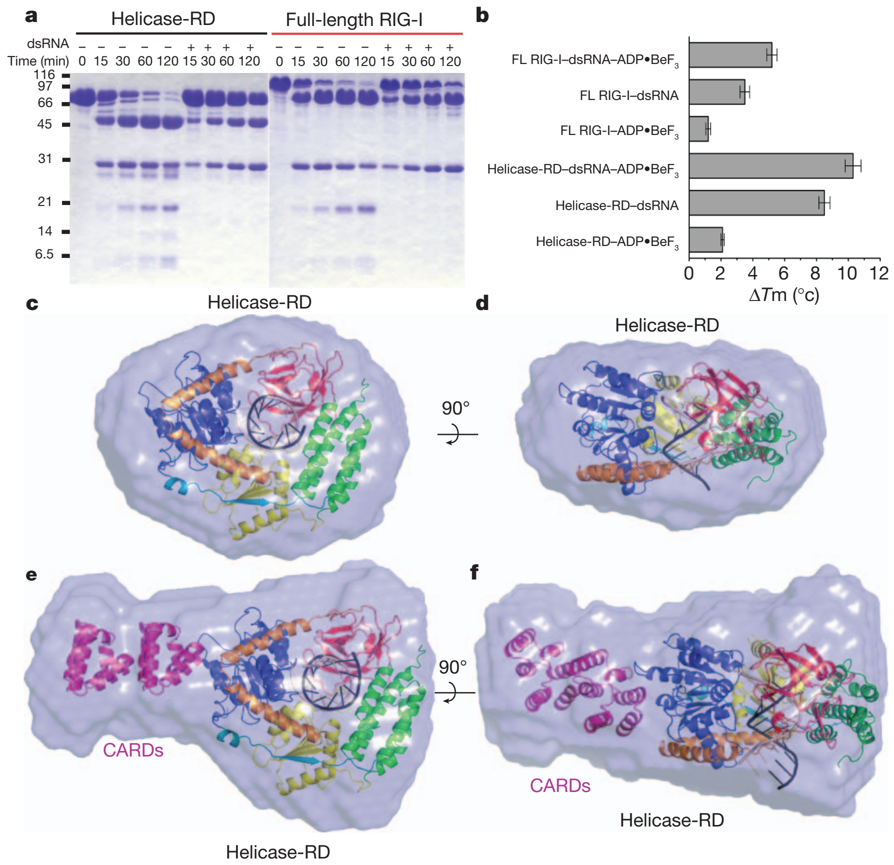


Figure 4 | Limited trypsin digestion, DSF and SAXS analyses of helicase-RD and full-length RIG-I in the presence and absence of dsRNA. **a**, SDS–PAGE analysis of a time course (minutes) of limited trypsin digestion of helicase-RD or full-length RIG-I in the absence or presence of 14 base-pair palindromic dsRNA. **b**, DSF of RIG-I helicase-RD or full-length (FL) RIG-I in the presence of 14 base-pair palindromic dsRNA and/or ADP•BeF_3 with respect to protein alone. The bar graph shows the mean melting temperature difference (ΔT_m) and the error bars represent the standard deviation from three independent measurements. **c**, **d**, *Ab initio* envelope of helicase-RD and dsRNA overlaid with the crystal structure of helicase-RD–dsRNA (dsRNA truncated to 10 base pairs). The view in **d** is rotated 90° about a horizontal axis from panel **c**. **e**, **f**, *Ab initio* envelope of full-length RIG-I and dsRNA overlaid with the crystal structure of helicase-RD–dsRNA with two copies of CARDs added (PDB accession 2VGQ). The view in **f** is rotated 90° about a horizontal axis from panel **e**.

RIG-I. Full-length RIG-I and helicase-RD undergo conformational changes and exhibit greater stability upon RNA binding (Fig. 4). The radius of gyration (R_g) of helicase-RD and full-length RIG-I decreases upon dsRNA binding by 10 and 2.3 Å, respectively (Supplementary Table 2), consistent with the results from size-exclusion chromatography (Supplementary Fig. 1b). The SAXS Kratky plots of helicase-RD and full-length RIG-I with RNA are symmetrical parabolic curves (Supplementary Fig. 5), consistent with folded and globular complexes. The parabolic shape of the Kratky plot is lost and the peak amplitude decreases in the absence of RNA, indicative of greater flexibility[19]. Limited trypsin digestion and DSF shows a greater stabilization with dsRNA and a small increase in stability upon addition of the ATP analogue (Fig. 4a, b). These data indicate that RIG-I is composed of globular domains connected with flexible linkers that become ordered upon RNA binding. Such conformational stabilization upon ATP and RNA binding has been documented in other helicases[20–22].

A model of full-length RIG-I bound to dsRNA was established from *ab initio* SAXS envelope followed by rigid body refinement of the helicase-RD–dsRNA complex and homologous CARD structures (Supplementary Table 2 and Fig. 4c–f). The helicase-RD–dsRNA complex and two copies of the CARD are accommodated in the full-length RIG-I SAXS envelope. The two CARDs project from domain 1. Such an orientation of the CARDs would allow for interaction with downstream signalling factors[23–28]. Interestingly, the second CARD is positioned adjacent to the V-shaped linker and close to the T770 phosphorylation site. The close proximity of the linker to the second CARD suggests a possible mechanism for the two α-helices to serve as a hinge for RD movement upon RNA binding, leading to RIG-I activation. The model indicates that the dsRNA extends along a perpendicular axis relative to the CARDs and hence can accommodate several RIG-I molecules[29].

METHODS SUMMARY

Recombinant full-length RIG-I (1–925), helicase-RD (232–925), helicase (232–794), RD (795–925) and selenomethione-derivatized helicase-RD were expressed in *Escherichia coli* and purified to homogeneity using immobilized metal ion affinity, hydroxyapatite and heparin affinity chromatography. Fluorescence anisotropy titrations were performed at 25 °C (ref. 30) ($\lambda_{excitation}$, 494 nm and $\lambda_{emission}$, 516 nm) using a fluorescein-labelled 14 base-pair dsRNA prepared by annealing 5′-GGAGAGAACCGCCU and 3′-CCUCUCUUGGCGGA-F RNA, where F is fluorescein. Crystals of the native and selenomethione helicase-RD with palindromic dsRNA (5′-CGACGCUAGCGUCG) and ADP•BeF_3•Mg^{2+} were obtained in 25% (w/v) PEG 3350, 0.25 M NaSCN, 100 mM MOPS (pH 7.8), 3% (v/v) 2,2,2-trifluoroethanol at 20 °C by hanging drop. The crystals belong to space group $P6_522$ with cell parameters $a = b = 174.9$ Å and $c = 110.9$ Å. The structure was determined by single-wavelength anomalous dispersion (SAD) to 3.2 Å resolution and refined against a 2.9 Å resolution native data set. The final model has an R_{work} and R_{free} of 0.199 and 0.287, respectively. SAXS measurements were performed on full-length RIG-I and helicase-RD in the absence and presence of dsRNA (5′-GCGCGCGCGC). Buffer subtraction and radius of gyration R_g were calculated from Guinier plots. The maximum particle size D_{max} was determined by scanning a range of values and comparing experimental scattered intensity $I(s)$ values to distance distribution function $P(r)$ transforms. Ten *ab initio* models were averaged and normalized spatial discrepancy (NSD) values were calculated. The helicase-RD–dsRNA and a homologous CARD structure were positioned into the *ab initio* model and χ^2 values were determined.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 July; accepted 6 September 2011.
Published online 25 September 2011.

1. Schlee, M. *et al.* Approaching the RNA ligand for RIG-I? *Immunol. Rev.* **227,** 66–74 (2009).
2. Saito, T., Owen, D. M., Jiang, F., Marcotrigiano, J. & Gale, M. Innate immunity induced by composition-dependent RIG-I recognition of hepatitis C virus RNA. *Nature* **454,** 523–527 (2008).
3. Uzri, D. & Gehrke, L. Nucleotide sequences and modifications that determine RIG-I/RNA binding and signaling activities. *J. Virol.* **83,** 4174–4184 (2009).
4. Matsumiya, T. & Stafforini, D. M. Function and regulation of retinoic acid-inducible gene-I. *Crit. Rev. Immunol.* **30,** 489–513 (2010).
5. Chattopadhyay, S. *et al.* Viral apoptosis is induced by IRF-3-mediated activation of Bax. *EMBO J.* **29,** 1762–1773 (2010).
6. Sun, Z., Ren, H., Liu, Y., Teeling, J. L. & Gu, J. Phosphorylation of RIG-I by casein kinase II inhibits its antiviral response. *J. Virol.* **85,** 1036–1047 (2011).
7. Malathi, K., Dong, B., Gale, M. Jr & Silverman, R. H. Small self-RNA generated by RNase L amplifies antiviral innate immunity. *Nature* **448,** 816–819 (2007).
8. Lu, C., Ranjith-Kumar, C. T., Hao, L., Kao, C. C. & Li, P. Crystal structure of RIG-I C-terminal domain bound to blunt-ended double-strand RNA without 5′ triphosphate. *Nucleic Acids Res.* **39,** 1565–1575 (2011).
9. Lu, C. *et al.* The structural basis of 5′ triphosphate double-stranded RNA recognition by RIG-I C-terminal domain. *Structure* **18,** 1032–1043 (2010).
10. Wang, Y. *et al.* Structural and functional insights into 5′-ppp RNA pattern recognition by the innate immune receptor RIG-I. *Nature Struct. Mol. Biol.* **17,** 781–787 (2010).
11. Singleton, M. R., Dillingham, M. S. & Wigley, D. B. Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.* **76,** 23–50 (2007).
12. Fairman-Williams, M. E., Guenther, U. P. & Jankowsky, E. SF1 and SF2 helicases: family matters. *Curr. Opin. Struct. Biol.* **20,** 313–324 (2010).
13. Dürr, H., Korner, C., Muller, M., Hickmann, V. & Hopfner, K. P. X-ray structures of the *Sulfolobus solfataricus* SWI2/SNF2 ATPase core and its complex with DNA. *Cell* **121,** 363–373 (2005).
14. Myong, S. *et al.* Cytosolic viral sensor RIG-I is a 5′-triphosphate-dependent translocase on double-stranded RNA. *Science* **323,** 1070–1074 (2009).
15. Takahasi, K. *et al.* Nonself RNA-sensing mechanism of RIG-I helicase and activation of antiviral immune responses. *Mol. Cell* **29,** 428–440 (2008).
16. Gu, M. & Rice, C. M. Three conformational snapshots of the hepatitis C virus NS3 helicase reveal a ratchet translocation mechanism. *Proc. Natl Acad. Sci. USA* **107,** 521–528 (2009).
17. Lam, A. M., Keeney, D. & Frick, D. N. Two novel conserved motifs in the hepatitis C virus NS3 protein critical for helicase action. *J. Biol. Chem.* **278,** 44514–44524 (2003).
18. Saito, T. *et al.* Regulation of innate antiviral defenses through a shared repressor domain in RIG-I and LGP2. *Proc. Natl Acad. Sci. USA* **104,** 582–587 (2007).
19. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40,** 191–285 (2007).
20. Lorsch, J. R. & Herschlag, D. The DEAD box protein eIF4A. 2. A cycle of nucleotide and RNA-dependent conformational changes. *Biochemistry* **37,** 2194–2206 (1998).
21. Polach, K. J. & Uhlenbeck, O. C. Cooperative binding of ATP and RNA substrates to the DEAD/H protein DbpA. *Biochemistry* **41,** 3693–3702 (2002).
22. Theissen, B., Karow, A. R., Kohler, J., Gubaev, A. & Klostermeier, D. Cooperative binding of ATP and RNA induces a closed conformation in a DEAD box RNA helicase. *Proc. Natl Acad. Sci. USA* **105,** 548–553 (2008).
23. Loo, Y. M. *et al.* Viral and therapeutic control of IFN-β promoter stimulator 1 during hepatitis C virus infection. *Proc. Natl Acad. Sci. USA* **103,** 6001–6006 (2006).
24. Kawai, T. *et al.* IPS-1, an adaptor triggering RIG-I- and Mda5-mediated type I interferon induction. *Nature Immunol.* **6,** 981–988 (2005).
25. Meylan, E. *et al.* Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. *Nature* **437,** 1167–1172 (2005).
26. Seth, R. B., Sun, L., Ea, C. K. & Chen, Z. J. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-κB and IRF 3. *Cell* **122,** 669–682 (2005).
27. Xu, L. G. *et al.* VISA is an adapter protein required for virus-triggered IFN-β signaling. *Mol. Cell* **19,** 727–740 (2005).
28. Gack, M. U. *et al.* TRIM25 RING-finger E3 ubiquitin ligase is essential for RIG-I-mediated antiviral activity. *Nature* **446,** 916–920 (2007).
29. Binder, M. *et al.* Molecular mechanism of signal perception and integration by the innate immune sensor retinoic acid inducible gene-I (RIG-I). *J. Biol. Chem.*, (2011).
30. Tang, G. Q., Bandwar, R. P. & Patel, S. S. Extended upstream A-T sequence increases T7 promoter strength. *J. Biol. Chem.* **280,** 40707–40713 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge access to beamlines X29 at the NSLS (National Synchrotron Light Source), LRL-CAT at APS (Advanced Photon Source), and G1 and F1 at CHESS (Cornell High Energy Synchrotron Source) and thank the NSLS, APS and CHESS staff. NSLS and APS are supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-98CH10886 and DE-AC02-06CH11357, respectively. CHESS is supported by the NSF and NIH/NIGMS through NSF award DMR-0936384, and the MacCHESS resource is supported by NIH/NCRR award RR-01646. Use of the LRL-CAT beamline facilities at Sector 31 was provided by Eli Lilly & Company. We would like to thank V. Rajagopal for initiating the biochemical experiments and guiding the project in the early stages. We thank E. Arnold, H. Berman, S. K. Burley, R. Gillilian, L. Morisco, W. Olson, T. Saito, A. Shatkin, A. Stock, H. Yang and M. Zhuravieva for providing helpful comments and assistance. This work was supported by NIH grants GM55310 to S.S.P. and AI080659 to J.M.

Author Contributions The project was initiated by M.G., J.M. and S.S.P. J.M. and S.S.P. designed and supervised the project. M.G. provided reagents and consultation. F.J. designed protein constructs and established purification protocols. A.R. generated all RNA reagents. F.J. and A.R. purified the complex and set up crystallization screens. F.J. optimized the crystal for data collection. J.M., M.T.M., F.J. and A.R. collected, processed and analysed the X-ray crystallographic data. M.T.M., F.J. and J.M. collected and analysed the SAXS data. A.R., G.-Q.T. and S.S.P. collected and analysed the RNA binding and ATPase assays. F.J. performed limited proteolysis and thermal melting assay. S.S.P. and J.M. wrote the paper and all authors contributed to editing.

Author Information Atomic coordinates have been deposited in the Protein Data Bank under accession code 3TMI. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.M. (jmarco@cabm.rutgers.edu) or S.S.P. (patelss@umdnj.edu).

METHODS

Protein expression and purification. All the protein constructs were expressed with an N-terminal 6×His–SUMO fusion. Human full-length RIG-I (1–925), helicase-RD (232–925) and helicase (232–794) were expressed in *Escherichia coli* strain Rosetta (DE3) (Novagen) as soluble proteins. The soluble fraction of helicase-RD was purified from the cell lysate using a Ni^{2+}-nitrilotriacetate (Qiagen) column. The recovered protein was then digested with Ulp1 protease to remove the 6×His–SUMO tag and further purified by hydroxyapatite column (CHT-II, Bio-Rad) and heparin sepharose column (GE Healthcare). Finally, purified helicase-RD was dialysed against 50 mM HEPES (pH 7.5), 50 mM NaCl, 5 mM DTT, 10% glycerol overnight at 4 °C, snap frozen in liquid nitrogen, and stored at −80 °C. The RIG-I RD (795–925) was expressed in *E. coli* BL21 Star (DE3) cells and the soluble fraction was purified to homogeneity using a Ni^{2+}-nitrilotriacetate column, cation exchange (HiTrap SP, GE Healthcare) and gel filtration chromatography. Selenomethionine (SeMet)-labelled helicase-RD was produced in Rosetta (DE3) cells grown in M9 minimal medium supplemented with 50 mg ml^{-1} L-SeMet (Sigma) and specific amino acids to inhibit endogenous methionine synthesis. The SeMet protein was purified using the same procedure as the unmodified protein.

Fluorescence anisotropy titrations. Fluorescence anisotropy measurements[31] were carried out on a Fluoro-Max-4 spectrofluorimeter (Horiba Jobin Yvon). Fluorescein-labelled 14 base-pair dsRNA (Dharmacon) (10 nM) was titrated with full-length RIG-I, helicase domain, RD or helicase-RD in 50 mM Tris-acetate (pH 7.5), 100 mM Na-acetate, 10 mM Mg-acetate, 5% glycerol, 5 mM DTT, 0.05% Tween 20 buffer. 5′-ppp dsRNA (GGAGAGAACCGCCU) was transcribed using T7 RNA polymerase, PAGE purified, and annealed to a complementary, fluorescein-labelled RNA. Fluorescein anisotropy was measured at 25 °C with excitation at 494 nm and emission at 516 nm. The helicase-RD titrations were also performed at lower dsRNA concentrations (1 and 2 nM) to obtain an accurate K_d. The observed fluorescence anisotropy (r_{obs}) as a function of protein concentration (P_t) was fit to equations (1) and (2) to obtain the equilibrium dissociation constant, K_d.

$$r_{obs} = r_b f_b + r_f(1 - f_b) \quad (1)$$

where r_f and r_b are the anisotropy values of free RNA and of the complex, f_b is the fraction of RNA bound in the protein–RNA complex and $f_b = [PR]/[R_t]$ (PR is the concentration of the protein–RNA complex and R_t is the total RNA concentration).

$$[PR] = \frac{(K_d + [P_t] + [R_t]) - \sqrt{(K_d + [P_t] + [R_t])^2 - 4[P_t][R_t])}}{2} \quad (2)$$

Initial anisotropy of the free fluorescein-labelled dsRNA was 0.10 for all experiments. The measurements and errors are from two independent experiments in Table 1.

ATPase activity. A time course (0–30 min) of the ATPase reaction was performed using RIG-I helicase-RD (5 nM), ATP (1 mM) spiked with [γ-^{32}P]ATP with or without 14 base-pair dsRNA (80 nM) in buffer containing 50 mM MOPS (pH 7.4), 5 mM $MgCl_2$, 5 mM DTT, 0.01% Tween 20 at 37 °C. The ATPase activity of the RIG-I helicase domain was measured using a higher amount of protein (100 nM) and RNA (1 µM). The quenched reactions were analysed by PEI-Cellulose-F TLC (Merck) developed in 0.4 M potassium phosphate buffer (pH 3.4). The TLC plates were exposed to a phosphorimager plate, analysed on a Typhoon phosphor-imager, and quantified using ImageQuant software. The ATPase rate was determined from the plots of [Pi] produced versus time and the rate constant values were calculated by dividing the ATPase rate by the respective enzyme concentration. The mean rate constant from two independent measurements and range is shown in Table 1.

Analytical size-exclusion chromatography. Analytical size-exclusion chromatography was carried out on an AKTA FPLC system (GE Healthcare). Proteins were loaded on to Superdex 200 10/300 GL column (GE Healthcare) equilibrated with a buffer containing 50 mM HEPES (pH 7.5), 50 mM NaCl, 5 mM $MgCl_2$, 5 mM DTT. The eluate was monitored by ultraviolet absorbance at 280 nm. For the complex, helicase-RD was incubated with 14 base-pair dsRNA at the molar ratio of 1:1.2 on ice for 15 min before the sample was applied onto the column.

Preparation of helicase-RD–dsRNA. The palindromic RNA oligonucleotide (5′-CGACGCUAGCGUCG-3′) (Dharmacon) was deprotected and desalted into 20 mM potassium phosphate (pH 7.4) before use. The dsRNA was prepared by incubating the RNA at 95 °C for 1 min followed by gradual cooling to 4 °C. The resulting dsRNAs were mixed with purified helicase-RD in a RNA:protein molar ratio of 1.2:1, incubated at room temperature (20 °C) for 15 min, and then purified by size-exclusion chromatography (Superdex200, GE Healthcare) with an elution buffer of 50 mM HEPES (pH 7.5), 50 mM NaCl, 5 mM DTT, 5 mM $MgCl_2$.

Crystallization and X-ray diffraction data collection. Helicase-RD–dsRNA–ADP•BeF_3 ternary complex was reconstituted by incubating 0.17 mM helicase-RD with 0.17 mM dsRNA, 2 mM ADP, 2 mM $BeCl_2$ and 10 mM NaF on ice for 30 min before crystallization. Crystals of native and SeMet-substituted complex were grown by the hanging-drop vapour diffusion method at 20 °C. Aliquots (2.5 µl) of 15 mg ml^{-1} of helicase-RD–dsRNA–ADP•BeF_3 complex in 50 mM HEPES (pH 7.5), 50 mM NaCl, 5 mM DTT, 5 mM $MgCl_2$ were mixed with 2.5 µl of reservoir solution containing 25% (w/v) PEG 3350, 0.25 M NaSCN, 100 mM MOPS (pH 7.8), 3% (v/v) 2,2,2-trifluoroethanol. Crystals appeared after 2–3 days, and they grew to a maximum size of 0.15 × 0.15 × 0.5 mm over the course of 8 days. 5-iodo-uridine derivative helicase-RD–dsRNA–ADP•BeF_3 crystals were grown under identical conditions. For cryogenic data collection, crystals were transferred into crystallization solutions containing 5% (v/v) (2R,3R)-(-)-2,3-butanediol as cryoprotectant and then flash-cooled at 100 K. SeMet SAD data set was collected at the X29A beamline of the NSLS. Native data set was recorded at the LRL-CAT 31-ID beamline of the APS. Data from iodouridine derivative crystals were collected at the CHESS F1 beamline. All diffraction data were integrated using iMosflm and scaled in SCALA[32].

Structure determination and refinement. Phases were determined using the SAD method. SHELXD/HKL2MAP[33] detected a total of 12 out of 14 possible selenium sites in the asymmetric unit. Initial phases were calculated with Phaser[32] and followed by density modification by DM[32]. Phase extension to 3.2 Å by SOLOMON[32] produced an electron-density map into which most of the protein and RNA residues could be built unambiguously. The model was built using Coot[34] and refined in PHENIX[35]. The final model, comprising ADP•BeF_3, 12 base-pair dsRNA, RIG-I residues 240–495, 504–522, 527–687 and 690–923 has R_{work} and R_{free} values of 0.199 and 0.287, respectively. Model validation demonstrated no outliers and 90% of the residues located the most favourable region of the Ramachandran plot[35]. Statistics of the data processing and structure refinement are summarized in Supplementary Table 1.

SAXS and structural modelling. SAXS data were collected at the CHESS beamline G1 using a Finger Lakes CCD X-ray detector with a sample-to-detector distance of 1,450 mm, covering the range of scattering vectors $0.01 < s < 0.233$ Å^{-1}, where $s = 4\pi\sin\theta/\lambda$ (2θ is the scattering angle and $\lambda = 1.296$ Å). All samples were buffer exchanged into 50 mM HEPES (pH 7.5), 50 mM NaCl, 5 mM DTT, 5% glycerol by size-exclusion chromatography (Hiload 16/26 Superdex200, GE Healthcare) to minimize the discrepancies in background subtraction. SAXS data were reduced using Data Squeeze. Various programs in ATSAS software package were used to process and evaluate the scattering data. Radius of gyration (R_g) was analysed using the Guinier approximation with low angle data ($s < 1.3/R_g$) using PRIMUS[36]. The residuals from the Guinier plots did not show signs of protein aggregation. The probability distribution of distances between scattering atoms within the macromolecule, P(r), and the maximum atom pair distance, D_{max}, were determined from the scattering data using GNOM[37]. The program DAMMIF[38] was used to calculate low-resolution *ab initio* reconstructions from experimental SAXS profiles. Ten to twenty independent models were aligned, filtered and averaged based on the occupancy using SUPCOMB and DAMAVER[39] to reconstruct the final *ab initio* envelope, as judged by averaged normalized spatial discrepancies (NSD) less than 1.0. The structure modelling was refined against the solution scattering data by rigid body docking using SASREF program[40]. The orientation and position of individual domains in the structure models were further manually adjusted to minimize the discrepancies (χ^2) between the calculated scattering intensities and experimental scattering intensities computed using program CRYSOL[41]. Kratky plots were calculated using Origin (OriginLab).

DSF. Thermal shift assay was conducted with 10 µM of RIG-I helicase-RD or full-length RIG-I with or without 12 µM of 14 base-pair palindromic dsRNA and/or 2 mM ADP•BeF_3 in 50 mM HEPES (pH 7.5), 50 mM NaCl, 5 mM $MgCl_2$, 5 mM DTT and a 5× dilution of SYPRO Orange dye (Invitrogen) as described[42]. The fluorescence signal as a function of temperature was recorded using a Real Time PCR machine (Applied Biosystems). The temperature gradient is performed in the range of 25–80 °C with a ramp of 0.2 °C over the course of 60 min. Control assays were carried out with buffer in the presence or absence of RNA. Data were analysed with the Excel-based worksheet DSF analysis, and the Boltzmann model was used to fit the fluorescence data to obtain the midpoint temperature for the thermal protein unfolding transition (T_m) using the curve-fitting software Prism.

Limited trypsin proteolysis. Limited proteolysis with trypsin (Roche) was performed using 120 µg of purified full-length RIG-I or helicase-RD in the absence or presence of palindromic dsRNA and incubated with trypsin at a protein:protease mass ratio of 300:1. The reaction mixtures were maintained at room temperature

and aliquots were removed at 15, 30, 60 and 120 min. The reaction was stopped by the addition of SDS–PAGE loading buffer and analysed by SDS–PAGE.

31. Tang, G. Q., Bandwar, R. P. & Patel, S. S. Extended upstream A-T sequence increases T7 promoter strength. *J. Biol. Chem.* **280,** 40707–40713 (2005).
32. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67,** 235–242 (2011).
33. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta Crystallogr. D* **58,** 1772–1779 (2002).
34. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
35. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D* **66,** 213–221 (2010).
36. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. & Svergun, D. I. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Cryst.* **36,** 1277–1282 (2003).
37. Semenyuk, A. V. & Svergun, D. I. GNOM: a program package for small-angle scattering data processing. *J. Appl. Cryst.* **24,** 537–540 (1991).
38. Svergun, D. I. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76,** 2879–2886 (1999).
39. Volkov, V. V. & Svergun, D. I. Uniqueness of ab initio shape determination in small-angle scattering. *J. Appl. Cryst.* **36,** 860–864 (2003).
40. Petoukhov, M. V. & Svergun, D. I. Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* **89,** 1237–1250 (2005).
41. Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **129,** 5656–5664 (2007).
42. Niesen, F. H., Berglund, H. & Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols* **2,** 2212–2221 (2007).

doi:10.1038/nature10506

Temperature-scan cryocrystallography reveals reaction intermediates in bacteriophytochrome

Xiaojing Yang[1], Zhong Ren[2], Jane Kuk[1] & Keith Moffat[1,2,3]

Light is a fundamental signal that regulates important physiological processes such as development and circadian rhythm in living organisms. Phytochromes form a major family of photoreceptors responsible for red light perception in plants, fungi and bacteria[1]. They undergo reversible photoconversion between red-absorbing (Pr) and far-red-absorbing (Pfr) states, thereby ultimately converting a light signal into a distinct biological signal that mediates subsequent cellular responses[2]. Several structures of microbial phytochromes have been determined in their dark-adapted Pr or Pfr states[3–7]. However, the structural nature of initial photochemical events has not been characterized by crystallography. Here we report the crystal structures of three intermediates in the photoreaction of *Pseudomonas aeruginosa* bacteriophytochrome (PaBphP). We used cryotrapping crystallography to capture intermediates, and followed structural changes by scanning the temperature at which the photoreaction proceeded. Light-induced conformational changes in PaBphP originate in ring D of the biliverdin (BV) chromophore, and *E*-to-*Z* isomerization about the $C_{15}{=}C_{16}$ double bond between rings C and D is the initial photochemical event. As the chromophore relaxes, the twist of the C_{15} methine bridge about its two dihedral angles is reversed. Structural changes extend further to rings B and A, and to the surrounding protein regions. These data indicate that absorption of a photon by the Pfr state of PaBphP converts a light signal into a structural signal via twisting and untwisting of the methine bridges in the linear tetrapyrrole within the confined protein cavity.

Cryotrapping and time-resolved room-temperature experiments are two main experimental strategies to study the structures of intrinsically short-lived reaction intermediates[8]. To establish the molecular mechanism of Pfr/Pr photoconversion, we generated and cryotrapped intermediates between the reactant Pfr state and product state(s) in fully photoactive crystals of the photosensory core module (PCM)[5] of *P. aeruginosa* BphP. We followed the progress of the reaction by applying a 'trap–pump–trap–probe' strategy at variable pump temperatures (Fig. 1 and Supplementary Fig. 1; Methods Summary). Here, temperature mimics time: the higher the pump temperature, the greater the structural relaxation and the further a reaction proceeds along its pathway.

We collected diffraction data from six crystals at ten pump temperatures between 100 and 180 K (Supplementary Table 1), and calculated difference ($F_{light} - F_{dark}$) electron density maps from 14 light data sets and 6 reference dark data sets (Supplementary Fig. 1a). In all maps, strong and highly significant difference electron densities are exclusively concentrated at the chromophore-binding sites embedded in the GAF domain (Fig. 1b). Singular value decomposition (SVD) analysis[9] of difference densities within a 5-Å radius of the aligned chromophores revealed three significant singular values, indicating three major, independent, light-induced structures (Supplementary Fig. 1c, d). Because difference densities are largely consistent among the eight monomers in the asymmetric unit and between different crystals illuminated at the same temperature, we averaged these densities by applying non-crystallographic symmetry (NCS), and focus here on the principal features common to the eight monomers.

Difference densities vary markedly with pump temperature (Fig. 2). At the lowest temperatures they appear near ring D of the BV chromophore, which suggests that light-induced structural changes originate in ring D. As the temperature rises, difference densities expand to ring C and eventually to ring B and ring A, thus exhibiting a systematic structural progression as a function of temperature. On the basis of representative difference maps at 110, 130 and 173 K, we modelled three light-induced chromophore structures, denoted L1, L2 and L3, respectively (Fig. 3a, b). We further refined the initial models of the L1, L2 and L3 structures jointly in real space against the NCS-averaged, SVD-filtered difference maps at all pump temperatures, and determined their relative concentrations at each temperature (Fig. 3g).

In the L1 structure represented at 110 K, strong positive and negative densities are roughly aligned in the plane of ring D of the bilin chromophore (Fig. 3a). Negative densities span the pyrrole nitrogen of ring D and the side chain of a highly conserved Asp 194 from the

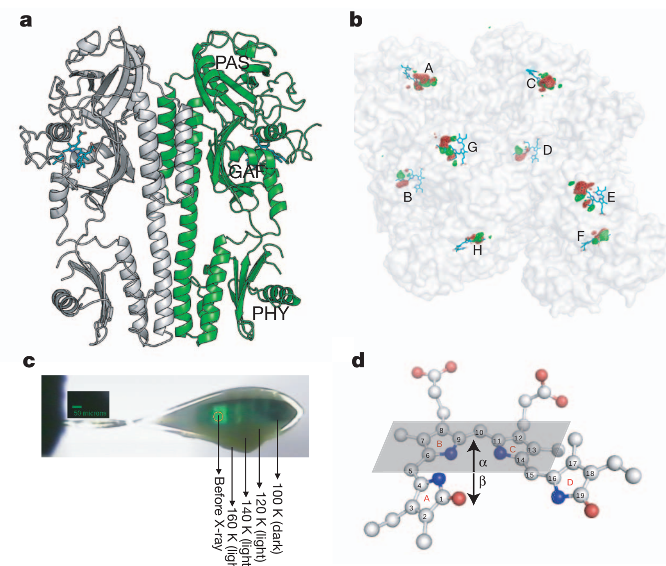


Figure 1 | Trap–pump–trap–probe experiment. a, Ribbon diagram of the PaBphP-PCM dimer. The BV chromophore is coloured in cyan. **b**, Experimental difference ($F_{light} - F_{dark}$) map at 130 K (contoured at $\pm 5\sigma$, where σ is the standard deviation of difference densities across the entire map). Strong positive (green) and negative (red) densities with peak signal greater than $\pm 12\sigma$ are clustered near the chromophores of the eight monomers (A–H) in the asymmetric unit. **c**, Dark stripes of a mounted crystal correspond to segments from which X-ray data sets were collected. **d**, A ball-and-stick representation of the chromophore in the Pfr state (PDB accession 3NHQ). The α-face of a pyrrole ring is defined when atom numbering follows a clockwise direction, with β defined as the opposite face. For definition of the α-face of an entire bilin chromophore see ref. 30.

[1]Department of Biochemistry and Molecular Biology, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, USA. [2]Center for Advanced Radiation Sources, The University of Chicago, 5610 South Ellis Avenue, Chicago, Illinois 60637, USA. [3]Institute for Biophysical Dynamics, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637, USA..

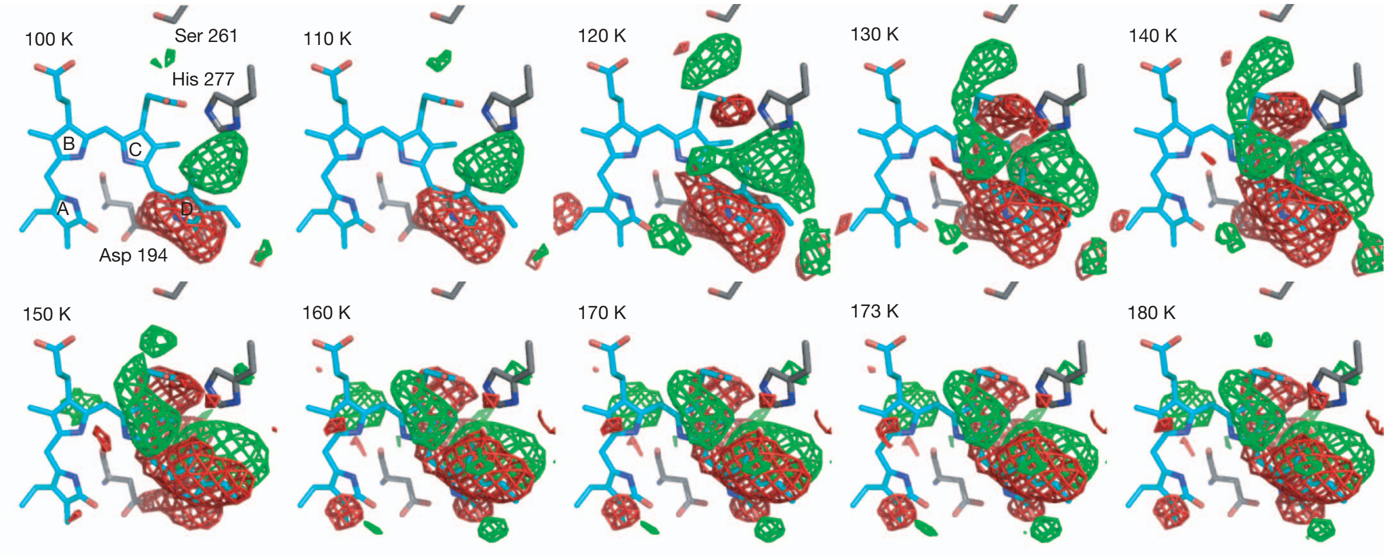
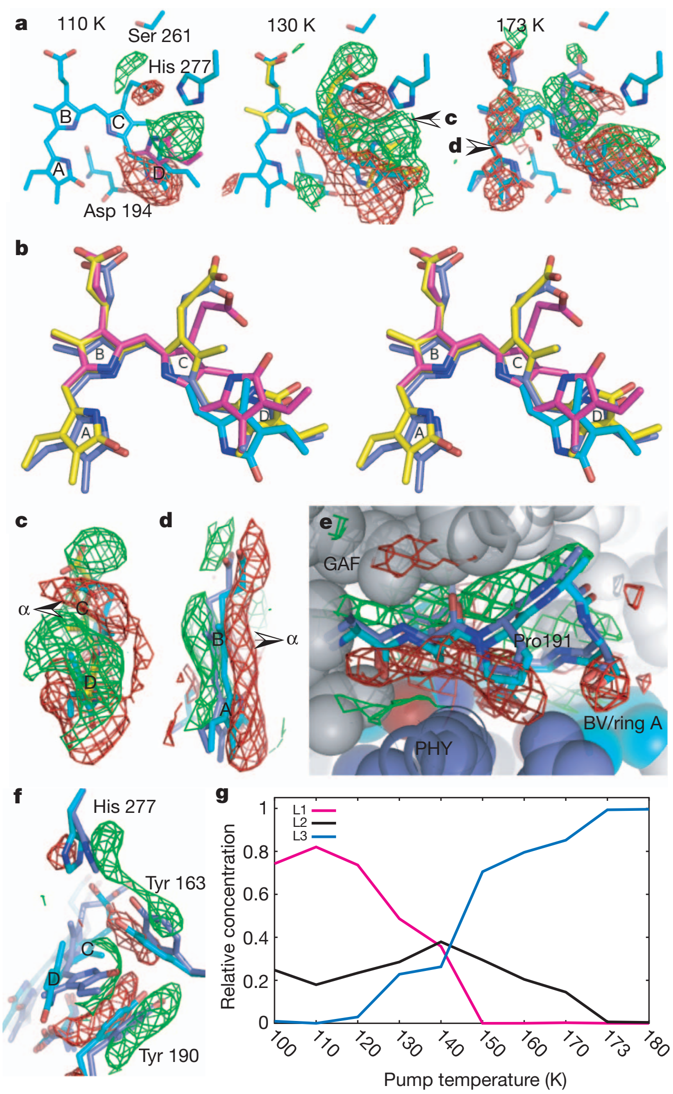


Figure 2 | Difference maps ($F_{light} - F_{dark}$) at pump temperatures between 100 and 180 K. The BV chromophore (cyan) is shown in the Pfr state. All maps are contoured at ±4σ. Positive (green) and negative (red) densities represent structural changes associated with formation of photoproduct state(s) and loss of the parent Pfr state, respectively. No significant difference densities with signal greater than ±4σ are detected at longer range, beyond the 5-Å radius around the chromophore.



PXSDIP sequence motif, indicating the rupture of a key hydrogen bond that stabilizes ring D in the Pfr state[5]. The corresponding positive densities between His 277 and ring D identify formation of the L1 structure in which the chromophore has isomerized about the $C_{15}=C_{16}$ double bond to adopt the 15*Za* configuration, and ring D is significantly shifted in its plane towards His 277. As a result, L1 exhibits a smaller and strained C_{14}–C_{15}–C_{16} bond angle in the methine bridge between rings C and D compared to the 'stretched' bond angle in the Pfr state (Supplementary Fig. 3) and to those of the Pr state in *Deinococcus radiodurans* BphP (DrBphP) and *Rhodopseudomonas palustris* BphP3 (RpBphP3) (refs 4, 10).

In the L2 structure represented at 130 K, strong difference densities near ring C suggest that ring C moves towards its α-face, and that its propionate side chain breaks the hydrogen bonds with His 277, Tyr 163 and Ser 275 present in the Pfr and L1 structures to form a new hydrogen bond with Ser 261 (Figs 3a and 4). Difference densities associated with rings C and D (Fig. 3c) suggest that counter-twist occurs across the C_{15} methine bridge in forming L2, in which ring D assumes a β-facial disposition relative to ring C. This partially relaxes strain in the C_{15} methine bridge while the C_{10} methine bridge between rings B and C is concomitantly twisted in the opposite direction (Supplementary Fig. 4b).

In the L3 structure represented at 173 K, structural changes extend to rings A and B (Fig. 3a). Difference densities sandwiching these rings (Fig. 3d) indicate that they move as a unit slightly towards the β-face of the chromophore while the C_5 methine bridge retains its direction of twist. Motion is accompanied by small twists in the C_5 and C_{10} methine

Figure 3 | Light-induced structural changes. **a**, Representative difference ($F_{light} - F_{dark}$) maps at 110, 130 and 173 K. Arrows indicate viewpoints of panels **c** and **d**. **b**, Stereo view of the superposition of the chromophore conformations in the Pfr (cyan), L1 (magenta), L2 (yellow) and L3 (blue) structures. **c**, Side view of difference map at 130 K (contoured at ±3σ) shows a twist in the C_{15} methine bridge. **d**, Side view of difference map at 173 K (contoured at ±2σ) indicates a β-facial shift of rings B/A. Arrows in **c** and **d** mark the α-face of the chromophore. **e**, Difference densities at 173 K (contoured at ±2σ) near the PXSDIP motif (residues 191–196) at the GAF–PHY interface. **f**, Differences densities associated with the side chains of His 277, Tyr 163 and Tyr 190 (map contours: ±3σ at 173 K). **g**, Relative concentrations of the L1, L2 and L3 structures as a function of pump temperature.

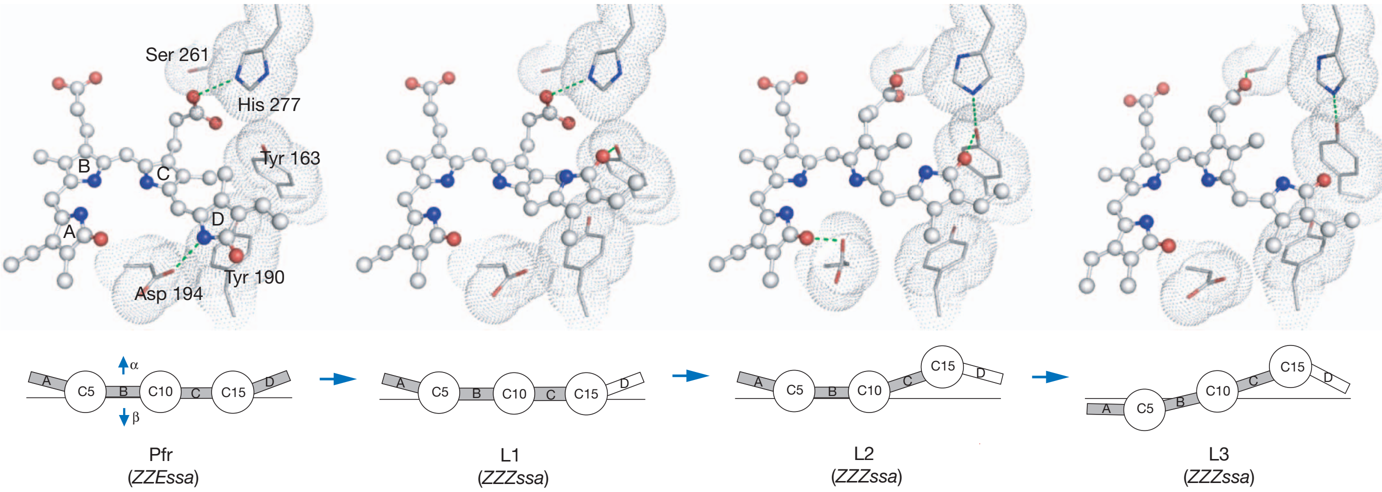


Figure 4 | Light-induced molecular events in PaBphP. Top, the chromophores are in ball-and-stick representation, with their surrounding residues shown as van der Waals spheres. Green dotted lines indicate potential interactions with each cryotrapped structure. Bottom, schematic representation of changes in relative disposition of the four pyrrole rings of the BV chromophore, in which pyrrole rings A, B, C and D (boxes) are linearly connected by methine bridges (circles). The α- and β-faces of the chromophore are denoted by arrows.

bridges (Supplementary Fig. 4b). Rings B and C are more coplanar in L3 than in L2, consistent with a more relaxed C_{10} methine bridge in L3.

These data demonstrate that *E*-to-*Z* isomerization about the $C_{15}{=}C_{16}$ double bond between rings C and D is the initial structural event in the Pfr-to-Pr photoreaction of PaBphP. This is consistent with findings in many phytochromes and bacteriophytochromes[2,11,12]; but contrasts with a recent report based on room temperature NMR spectroscopy, in which rotation of ring A occurs between the Pr and Pfr states of an unusual, knotless phytochrome consisting of only the GAF domain[13]. In PaBphP, when ring D flips into the 15*Za* configuration within a protein cavity that is still optimized to accommodate the 15*Ea* configuration of the Pfr state, a highly distorted and strained C_{15} methine bridge is generated in L1. As L1 evolves to L2, counter-twist across the C_{15} methine bridge partially relieves the strain. Motion of ring C results in a C_{10} methine bridge twisted in the opposite direction in L2, further relaxation of which leads to the L3 structure. These data indicate a reaction trajectory that proceeds in the order Pfr → L1 → L2 → L3 (Fig. 4).

To relate these cryotrapped L1, L2 and L3 structures to spectroscopic intermediates in the Pfr-to-Pr photoreaction[14–18], we measured visible absorption spectra on a crushed PaBphP-PCM crystal at temperatures between 100 and 180 K using the X-ray experimental protocol (Supplementary Fig. 5). Difference absorption spectra between illuminated and reference 'dark' states show significant loss of the Pfr state, indicated by a negative peak at 768 nm, and formation of blue-shifted photoproducts (Supplementary Fig. 5b). SVD analysis revealed two significant basis difference absorption spectra, with blue-shifted peaks between 650 and 700 nm. The basis spectrum with a blue-shifted peak at 684 nm is largely populated between 100 and 150 K, in agreement with the temperature dependence of the population of the L2 structure in X-ray data (Supplementary Fig. 5c and Fig. 3g). Blue-shifted absorption peaks are consistent with the distorted tetrapyrrole conjugated system observed in L2 and L3.

The L1, L2 and L3 structures probably arise from early molecular events in the Pfr-to-Pr reaction of PaBphP. One or more may correspond to the Lumi-F spectroscopic intermediate detected on the femto- to picosecond timescale by room-temperature time-resolved spectroscopy and characterized by blue-shifted absorption peaks[14,18,19]. These structures are consistent with strong difference bands from Fourier transform infrared (FTIR) spectroscopy that were attributed to the carbonyl group of ring D resulting from 15*Ea*-to-15*Za* isomerization and the B–C methine stretching in cryotrapped Lumi-F of *Calothrix* CphA[18]. L1 may be identified with early intermediates in *Synechocystis* Cph1, in which the C_{15}–H out-of-plane (HOOP) mode was detected by ultrafast Raman spectroscopy[20]. The properties of L1 are also consistent with the NMR spectroscopic data on the cryotrapped Lumi-F of Cph1, which indicated that the C_{14}–C_{15}–C_{16} angle is distorted following 15*Ea*-to-15*Za* isomerization[19]. Furthermore, our temperature-scanning range (100–180 K) coincides with the temperature range in which Lumi-F intermediates were trapped in several phytochrome systems[11,18,19]. Formation of late Meta-F intermediates would require higher pump temperatures[19] to permit more extensive structural relaxation in the protein. However, at temperatures >180 K the cryoprotectant solution undergoes a glass transition which causes severe deterioration in crystal diffraction[5,6].

In both the L2 and L3 structures, the twist of the dihedral angles in the C_{15} methine bridge is quite distinct from those of the chromophore structures in the Pr and Pfr states (Figs 3 and 4 and Supplementary Fig. 4b). Both the Pr and Pfr crystal structures of BphPs[3–5,10] exhibit an α-facial disposition of ring D relative to ring C, which corresponds to a negative rotation of the red absorbance band in the circular dichroism spectra of BV-containing bacteriophytochromes such as PaBphP, DrBphP and *Agrobacterium tumefaciens* Agp1 (refs 21, 22). Evidence for such counter-twist events in the methine bridges during photoconversion has been presented in Cph1 and Agp1 (refs 19, 22). We propose that the BV chromophore of PaBphP generates structural signals in response to light via subtle and local twist or counter-twist motions in the methine bridges within the confined protein cavity (Fig. 4), which alter specific interactions between the pyrrole rings and their immediate protein surroundings. It remains to be seen to what extent our findings in PaBphP apply to other members of the diverse and expanding phytochrome superfamily.

The protein moiety also has an important role. First, following prompt *E*-to-*Z* isomerization, steric clashes between ring D and the side chain of Asp 194 lead to initial strain in L1 that drives subsequent relaxation events. Second, the side chains of His 277, Tyr 163 and Tyr 190 surrounding ring D also move slightly but concertedly to accommodate ring D as the chromophore evolves from Pfr to the L3 structure (Fig. 3f). Third, Ser 261 seems to stabilize L2 and L3 via hydrogen bonds to the propionate group of ring C (Fig. 3a, 4). The single point mutant S261A inhibits formation of the Pr state upon illumination and significantly accelerates dark reversion[23] (Supplementary Fig. 6b). This provides further evidence that L2 and L3 are authentic photoproducts on the productive trajectory towards the product Pr state. Fourth, more extensive structural changes occur in L3 near the highly conserved PXSDIP sequence motif at the interface of

the GAF and PHY domains. This segment and the side chain of Tyr 190 are pushed away from the chromophore as rings B/A and ring D move towards the β-face of the chromophore (Fig. 3e). These movements expand the chromophore cavity to allow further relaxation, and may trigger further structural perturbations between the GAF domain and the arm of the PHY domain. As the reaction proceeds further, Tyr 163 and Tyr 190 may adjust their side-chain rotamers to reshape the ring D pocket to accommodate the product Pr state[23].

To examine the light dependence of histidine kinase (HK) activity in PaBphP, we conducted HK assays on wild type and the S261A mutant of full-length PaBphP under light and dark conditions (Supplementary Fig. 6) and found that autophosphorylation of PaBphP is light dependent, in contrast to earlier observations[24]. As in *R. palustris* BphP2 (ref. 25) and Cph1 (ref. 26), PaBphP shows higher levels of autophosphorylation in the Pr state (here, the light state) than in the Pfr state. S261A exhibits a photoconversion phenotype that greatly prefers the Pfr state, and shows significantly reduced HK activity compared to wild type. We predict that small light-induced structural changes, exemplified by the cryotrapped intermediate structures L1, L2 and L3, propagate from the chromophore-binding pocket of the amino-terminal photosensory domains to the carboxy-terminal HK domain via further tertiary and/or quaternary rearrangements[27], which ultimately affect the HK activity and convert a light signal into a biological signal.

METHODS SUMMARY

Structural heterogeneity is intrinsic to all dynamic processes, and often challenges accurate interpretation of both cryotrapping experiments at a single temperature and time-resolved crystallographic data[28,29]. This work presents a temperature-scan and analytical strategy to resolve structural heterogeneity and determine distinct, homogeneous structural species.

We applied a 'trap–pump–trap–probe' strategy to photoactive crystals of PaBphP-PCM. We first trapped the dark-adapted Pfr state (trap dark) by freezing the crystal (grown at 293 K in the dark) in liquid nitrogen under safety green/blue light, and collected a reference crystallographic 'dark' data set at 100 K for each crystal. The same crystal was then uniformly illuminated under white light for 10–15 min at elevated temperatures to generate reaction intermediates (pump), then cryo-cooled to 100 K (trap light) to collect a 'light' data set from a fresh crystal volume (probe) (Fig. 1c). We calculated 14 ($F_{\mathrm{light}} - F_{\mathrm{dark}}$) difference maps from six crystals that cover 10 pump temperatures between 100 and 180 K. As each map contains eight monomers in the asymmetric unit, we obtained 112 independent difference maps for a PaBphP monomer. We spatially aligned these maps on a reference monomer C, and subjected difference densities within a 5-Å radius of the aligned chromophores to singular value decomposition (SVD)[9] that identified three significant singular values and yielded 112 noise-reduced difference maps (Supplementary Fig. 1c, d). Atomic models for the L1, L2 and L3 structures were initially built based on representative difference densities at 110, 130 and 173 K, respectively. These models were further refined jointly in real space (together with a fixed reference Pfr conformation) against SVD-filtered difference maps at all pump temperatures. On the basis of least squares fitting between the calculated and observed difference densities, we also obtained relative concentrations of L1, L2 and L3 at each temperature. Supplementary Fig. 1a summarizes methods and software used in X-ray data reduction and analysis.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 June 2010; accepted 23 August 2011.
Published online 16 October 2011.

1. Montgomery, B. L. & Lagarias, J. C. Phytochrome ancestry: sensors of bilins and light. *Trends Plant Sci.* **7,** 357–366 (2002).
2. Rockwell, N. C., Su, Y. S. & Lagarias, J. C. Phytochrome structure and signaling mechanisms. *Annu. Rev. Plant Biol.* **57,** 837–858 (2006).
3. Wagner, J. R., Brunzelle, J. S., Forest, K. T. & Vierstra, R. D. A light-sensing knot revealed by the structure of the chromophore-binding domain of phytochrome. *Nature* **438,** 325–331 (2005).
4. Yang, X., Stojkovic, E. A., Kuk, J. & Moffat, K. Crystal structure of the chromophore binding domain of an unusual bacteriophytochrome, RpBphP3, reveals residues that modulate photoconversion. *Proc. Natl Acad. Sci. USA* **104,** 12571–12576 (2007).
5. Yang, X., Kuk, J. & Moffat, K. Crystal structure of *Pseudomonas aeruginosa* bacteriophytochrome: photoconversion and signal transduction. *Proc. Natl Acad. Sci. USA* **105,** 14715–14720 (2008).
6. Essen, L. O., Mailliet, J. & Hughes, J. The structure of a complete phytochrome sensory module in the Pr ground state. *Proc. Natl Acad. Sci. USA* **105,** 14709–14714 (2008).
7. Cornilescu, G., Ulijasz, A. T., Cornilescu, C. C., Markley, J. L. & Vierstra, R. D. Solution structure of a cyanobacterial phytochrome GAF domain in the red-light-absorbing ground state. *J. Mol. Biol.* **383,** 403–413 (2008).
8. Moffat, K. & Henderson, R. Freeze trapping of reaction intermediates. *Curr. Opin. Struct. Biol.* **5,** 656–663 (1995).
9. Rajagopal, S., Schmidt, M., Anderson, S., Ihee, H. & Moffat, K. Analysis of experimental time-resolved crystallographic data by singular value decomposition. *Acta Crystallogr. D* **60,** 860–871 (2004).
10. Wagner, J. R., Zhang, J., Brunzelle, J. S., Vierstra, R. D. & Forest, K. T. High resolution structure of *Deinococcus* bacteriophytochrome yields new insights into phytochrome architecture and evolution. *J. Biol. Chem.* **282,** 12298–12309 (2007).
11. Foerstendorf, H., Mummert, E., Schafer, E., Scheer, H. & Siebert, F. Fourier-transform infrared spectroscopy of phytochrome: difference spectra of the intermediates of the photoreactions. *Biochemistry* **35,** 10793–10799 (1996).
12. Song, C. *et al.* Two ground state isoforms and a chromophore *D*-ring photoflip triggering extensive intramolecular changes in a canonical phytochrome. *Proc. Natl Acad. Sci. USA* **108,** 3842–3847 (2011).
13. Ulijasz, A. T. *et al.* Structural basis for the photoconversion of a phytochrome to the activated Pfr form. *Nature* **463,** 250–254 (2010).
14. Schumann, C. *et al.* Subpicosecond midinfrared spectroscopy of the P_{fr} reaction of phytochrome Agp1 from *Agrobacterium tumefaciens*. *Biophys. J.* **94,** 3189–3197 (2008).
15. van Thor, J. J., Ronayne, K. L. & Towrie, M. Formation of the early photoproduct lumi-R of cyanobacterial phytochrome Cph1 observed by ultrafast mid-infrared spectroscopy. *J. Am. Chem. Soc.* **129,** 126–132 (2007).
16. van Wilderen, L. J., Clark, I. P., Towrie, M. & van Thor, J. J. Mid-infrared picosecond pump–dump–probe and pump–repump–probe experiments to resolve a ground-state intermediate in cyanobacterial phytochrome Cph1. *J. Phys. Chem. B* **113,** 16354–16364 (2009).
17. Muller, M. G., Lindner, I., Martin, I., Gartner, W. & Holzwarth, A. R. Femtosecond kinetics of photoconversion of the higher plant photoreceptor phytochrome carrying native and modified chromophores. *Biophys. J.* **94,** 4370–4382 (2008).
18. Schwinté, P. *et al.* The photoreactions of recombinant phytochrome CphA from the cyanobacterium *Calothrix* PCC7601: a low-temperature UV–Vis and FTIR study. *Photochem. Photobiol.* **85,** 239–249 (2009).
19. Rohmer, T. *et al.* Phytochrome as molecular machine: revealing chromophore action during the Pfr → Pr photoconversion by magic-angle spinning NMR spectroscopy. *J. Am. Chem. Soc.* **132,** 4431–4437 (2010).
20. Dasgupta, J., Frontiera, R. R., Taylor, K. C., Lagarias, J. C. & Mathies, R. A. Ultrafast excited-state isomerization in phytochrome revealed by femtosecond stimulated Raman spectroscopy. *Proc. Natl Acad. Sci. USA* **106,** 1784–1789 (2009).
21. Rockwell, N. C., Shang, L., Martin, S. S. & Lagarias, J. C. Distinct classes of red/far-red photochemistry within the phytochrome superfamily. *Proc. Natl Acad. Sci. USA* **106,** 6123–6127 (2009).
22. Seibeck, S. *et al.* Locked 5Zs-biliverdin blocks the Meta-R_A to Meta-R_C transition in the functional cycle of bacteriophytochrome Agp1. *FEBS Lett.* **581,** 5425–5429 (2007).
23. Yang, X., Kuk, J. & Moffat, K. Conformational differences between the Pfr and Pr states in *Pseudomonas aeruginosa* bacteriophytochrome. *Proc. Natl Acad. Sci. USA* **106,** 15639–15644 (2009).
24. Tasler, R., Moises, T. & Frankenberg-Dinkel, N. Biochemical and spectroscopic characterization of the bacterial phytochrome of *Pseudomonas aeruginosa*. *FEBS J.* **272,** 1927–1936 (2005).
25. Giraud, E. *et al.* A new type of bacteriophytochrome acts in tandem with a classical bacteriophytochrome to control the antennae synthesis in *Rhodopseudomonas palustris*. *J. Biol. Chem.* **280,** 32389–32397 (2005).
26. Yeh, K. C., Wu, S. H., Murphy, J. T. & Lagarias, J. C. A cyanobacterial phytochrome two-component light sensory system. *Science* **277,** 1505–1508 (1997).
27. Li, H., Zhang, J., Vierstra, R. D. & Li, H. Quaternary organization of a phytochrome dimer as revealed by cryoelectron microscopy. *Proc. Natl Acad. Sci. USA* **107,** 10872–10877 (2010).
28. Anderson, S., Srajer, V. & Moffat, K. Structural heterogeneity of cryotrapped intermediates in the bacterial blue light photoreceptor, photoactive yellow protein. *Photochem. Photobiol.* **80,** 7–14 (2004).
29. Schmidt, M. *et al.* Protein kinetics: structures of intermediates and reaction mechanism from time-resolved X-ray data. *Proc. Natl Acad. Sci. USA* **101,** 4799–4804 (2004).
30. Rockwell, N. C. *et al.* A second conserved GAF domain cysteine is required for the blue/green photoreversibility of cyanobacteriochrome Tlr0924 from *Thermosynechococcus elongatus*. *Biochemistry* **47,** 7304–7316 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Möglich for comments and reading of the manuscript, and V. Šrajer of BioCARS for assistance in microspectrometer experiments on crystals. We also thank the staff of LSCAT and BioCARS at the Advanced Photon Source, Argonne National Laboratory for beamline access. Supported by National Institutes of Health grant GM036452 to K.M. BioCARS is supported by National Institutes of Health grant RR07707 to K.M.

Author Contributions X.Y. initiated and designed research, collected X-ray and microspectroscopic data; carried out mutagenesis and HK assays; X.Y. and Z.R. analysed and interpreted structures; Z.R. developed data analysis methods and analysed data; J.K. purified proteins and grew crystals; K.M. initiated photoreceptor projects; X.Y., Z.R. and K.M. wrote the manuscript.

Author Information Atomic coordinates and structure factor amplitudes have been deposited in the Protein Data Bank under accession codes 3NHQ, 3NOP, 3NOT and 3NOU. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to X.Y. (xiaojingyang@uchicago.edu) or K.M. (moffat@cars.uchicago.edu).

METHODS

Purification, mutagenesis and crystallization. PaBphP-PCM containing residues 1–497 of *P. aeruginosa* PA4117 was expressed, purified, crystallized and cryo-protected in the dark as described[5]. Site-directed mutagenesis was carried out using the QuikChange Site-directed Mutagenesis Kit (Stratagene). Both wild type and the S261A mutant of the full-length PaBphP were purified using the same protocol as for PaBphP-PCM.

Data collection and analysis. Diffraction data were collected at the LSCAT 21-IDG beamline, BioCARS 14-BMC and 14-IDB beamlines of the Advanced Photon Source. Temperature was controlled by a cryostream cooler (Oxford Cryosystems). Illumination was applied to a mounted crystal using two unfiltered fibre optic lights from different directions for 10–15 min, while the crystal rotated around the spindle axis of the goniometer at $30^{\circ}\ s^{-1}$. All diffraction images were processed using HKL2000[31]. Light and dark data sets from the same crystal were scaled using ScaleIt in CCP4 (ref. 32; Supplementary Table 1). A reference 'dark' structure in the Pfr state was refined at 2.55-Å resolution using Phenix[33] (Supplementary Table 2), and was used to calculate phases in generating difference Fourier maps by FFT in CCP4. Difference maps were masked and aligned using NCSMASK and MAPROT of CCP4, and were further subjected to SVD[9] analysis in DynamiX (Supplementary Fig. 1). Real space refinement of the L1, L2 and L3 structures, including the chromophore and several adjacent residues (Gln 188–Asp 194, His 277, Ser 261, Tyr 163 and the covalent chromophore anchor Cys 12), was carried out against NCS-averaged, SVD-filtered difference maps using DynamiX (Supplementary Table 3). All modelling building was carried out using Coot[34]. Structure figures were generated using PyMol (http://pymol.org)

Ultraviolet-visible spectroscopy. Absorption spectra of proteins in solution were recorded at room temperature ($25 \pm 2\ ^{\circ}C$) with a Shimadzu UV-1650 PC spectrophotometer. Visible spectra of crystals were measured using a microspectrophotometer Xspectra (4DX-ray Systems) at BioCARS.

HK autophosphorylation assay. The HK assay was modified based on published protocols[26,35]. Reactions under light (continuous illumination with unfiltered fibre optical light) or dark (samples were protected from light in a covered box during dark incubation and reaction; handled under dim room light) were started by adding 10 µl reaction buffer (50 mM Tris HCl, pH 8.0; 2 mM $MgCl_2$; 2 mM $MnCl_2$; 0.1 M KCl; 10% ethylene glycol; 80 µM ATP and ^{32}P-γ-ATP at 0.3 mCi ml^{-1}) to each 10 µl protein sample at the concentration of 1 mg ml^{-1}. Reactions were stopped using standard stop buffer containing 0.1 M EDTA and 0.1 M dithiothreitol.

31. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
32. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67,** 235–242 (2011).
33. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
34. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
35. Giraud, E., Lavergne, J. & Vermeglio, A. Characterization of bacteriophytochrome from photosynthetic bacteria: histidine kinase signaling triggered by light and redox sensing. *Methods Enzymol.* **471,** 135–159 (2010).

CAREERS

NETWORKING In United States, student demand for mentors rises **p.435**

PHD CANDIDATES A change in student status could make research more attractive **p.435**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

OCEAN/CORBIS

A roll of the dice

For some, a lack of tenure creates a dynamic lab environment. For others, it's a gamble not worth taking.

BY KAREN KAPLAN

Sean Eddy has his dream job: he is a group leader in computational genomics at the Janelia Farm Research Campus of the Howard Hughes Medical Institute (HHMI), in Ashburn, Virginia. Yet, as he approaches his first cyclical review next year, he faces the prospect of being asked to leave if his work is not deemed worthy of the institute's mission.

Eddy was one of ten scientists who, aiming to energize their research and forge multidisciplinary ties, decided in 2006 to join a newly opened research institute with unconventional operating and funding models. Although he was once a tenured researcher at the Washington University School of Medicine in St Louis, Missouri, Eddy is unruffled by the lack of tenure at Janelia. In July 2012, Eddy will undergo a review, required for all Janelia group leaders — there are now 26 — after their initial six years. If an external review panel finds his work deserving, he will be offered the chance to renew for five years. If his work doesn't measure up, he has to be out by July 2014. But the uncertainty of his future never keeps him up at night.

Eddy is "wonderfully stressed" about the review. "I like knowing they can kick me out to the street," he says. But he isn't revealing a masochistic streak. Remove the security blanket of tenure, says Eddy, and he is driven to work harder, and to assess his research programme more frequently to make sure that it is still on the right track. Furthermore, he says, tenure, which is especially coveted in the United States, brings its own job-related anxieties. "If I'm tenured at Washington University or anywhere, they can't fire me, but they can put me in a closet and take away my space," he says. "I prefer it this way — I think it's appropriate to have a little fire under you."

TOP MODEL?

As Janelia reaches its fifth anniversary, its research and culture continue to draw notice, and the question of whether its approach is effective remains unanswered (see page 284). Its operating model was a head-turner in 2000, when the HHMI announced plans to create the research campus; and when Janelia opened in 2006, it sparked articles in the academic, scientific and mainstream press that noted its 'radical' departure from the conventional US academic approach (see *Nature* **443,** 128–129; 2006).

But executive director Gerry Rubin, a former academic, emphasizes that Janelia's cyclical-review model is not new. It is based in large part on similar models at established institutes that offer fixed-term contracts with reviews and opportunities to renew, such as the Medical Research Council Laboratory of Molecular Biology (LMB) in Cambridge, UK, and the former basic-research model at Bell Laboratories in Murray Hill, New Jersey, which is now the research arm of French telecommunications company Alcatel-Lucent. Similar models at Cold Spring Harbor Laboratory, a biological sciences institute in New York, and the Carnegie Institution for Science, based in Washington DC, also helped to inspire Janelia. The European Molecular Biology Laboratory, which has five sites across Europe, offers rotating contracts too (see *Nature* **478,** 547–548; 2011).

Scientists at Janelia and similar institutions don't baulk at giving up the comfort and protection of a longer-term job — and in many cases, tenure. On the contrary, they're eager to abandon the academic prototype in favour of a workplace culture in which research is the ▶

focus and high-risk, inventive projects are the norm. They are also generally less worried about grants, teaching, committee service and other off-the-bench activities. Indeed, despite the job security and intellectual freedom that tenure confers, it is hardly universally relevant or obligatory, argue administrators and some bench scientists. Limited-term, research-focused contracts, they say, sharpen research programmes by ensuring that scientists are actively involved in day-to-day experiments.

Still, only researchers with an appetite for high-risk work and a willingness to change institutions and lab environments should embrace such a model. Young scientists should also keep in mind that labs at these institutions tend to be far smaller than those in academia, which could create logistical problems if people leave. Researchers who enjoy teaching or the university setting are also more likely to find career satisfaction elsewhere.

At Janelia Farm Research Campus, scientists forgo tenure for short-term contracts and cutting-edge labs.

M. STALEY/JANELIA FARM

TENURE TIME-OUT

From the start, Rubin felt sure that Janelia held promise. "We looked at the LMB and Bell and Cold Spring and Carnegie and we saw that you did not have to offer tenure to get the highest quality of scientists," he says.

Tenure can be antithetical to good science, says Eric Betzig, a group leader in physics at Janelia, who spent six years at Bell. "The chase for tenure enforces a certain conservatism — you learn not to stick your neck out," says Betzig. "Then, once you have it, it's possible to get stale. And it's small enough around here that we can't afford to have a bunch of stale people."

Limited-contract institutions typically provide generous funding packages, with a salary for four to five years and enough money to buy equipment and supplies, and hire a postdoc and lab technician. The publish-or-perish imperative of academia is greatly reduced, because such institutions focus more on the researcher's overall scientific programme than on his or her publication rate.

And, because few of these institutions, at least in the United States, offer classes for students, scientists working at them typically don't have to teach; instead, researchers spend a lot of time in the lab. At some facilities, such as Janelia and Bell, scientists have virtually no obligations outside their research; Janelia, in fact, requires its scientists to spend 75% of their time at the bench. Other organizations require a nominal level of non-research commitment, such as service on a committee. "The postdocs here are ticked off because the principal investigators are having so much fun," says Eddy. "At Janelia, we're all saying, 'Yeah, I guess I should let the postdoc do an experiment'." Harald Hess, a group leader doing high-resolution microscopy at Janelia, who also spent 11 years at Bell, says that there are few time-sinks to keep scientists away from the bench at either institution. Rubin agrees. "If you want to work in the lab with your own hands, you have to come here," he says. "That's not going to happen at most academic institutions."

In return for the right to concentrate so closely on their research, scientists tend to be reviewed on how innovative their programmes are, and on the likelihood of field-changing discoveries, rather than on more conventional metrics. "You may not succeed, and you may not have anything to show for your five or seven years," says Karel Svoboda, a neurobiologist and biophysicist at Janelia who has worked at both Bell and Cold Spring Harbor. "But in this environment, you may still be viewed as successful, even if you don't have the big paper."

JUDGEMENT DAY

Review committees at non-tenure institutions examine investigators' work at set intervals, usually every four or five years; researchers who don't make the cut generally have between six months and two years to find a new position. Panels can be internal, external or a combination of both. For example, when the first reviews start happening at Janelia, the committee will consist of about 20 scientists, half from the group that reviews HHMI-funded investigators at other institutions, and half from the field of the person being reviewed. The reviewees will give 45-minute presentations on their work to the full panel.

Review criteria vary, but institutions strive to ensure that their researchers' science is original and creative, and will have an impact. "We don't just count papers or citations, we make a judgement about whether people are doing something that's worth doing," says Hugh Pelham, director of the LMB. Carnegie asks whether the reviewees are taking advantage of the opportunities provided by the institution, notes president Richard Meserve — in particular, that they are effectively using the time freed up by not having to teach or chase grants. Institutions may consider how much collaboration principal investigators have been involved in and how active they have been on committees; Rubin says he will also provide input on reviewees' performance as lab colleagues and mentors to junior scientists. At the LMB, Pelham and others who regularly interact with reviewees can step in and disagree with the panel's comments; Pelham can even override a recommendation to dismiss, if he thinks the reviewee is on the cusp of a big breakthrough.

At Janelia, investigators aren't allowed to seek external funding, so grant success is irrelevant in reviews. But this is not true everywhere: for example, Cold Spring Harbor does take grant success, and indeed publication rate, into account. Its internal review panel uses both to gauge whether investigators have developed independent research programmes and have the potential to become leaders in their fields. Ideally, the lab would like investigators who are renewed in their fourth-year reviews to earn enough external funding to support 80% of their work by their fifth year.

Meserve declines to reveal Carnegie's staff-retention rates, but says that "very few" of the scientists hired as permanent staff members have left in the past two cycles. Rubin expects about 80% retention at Janelia.

RISKY BUSINESS

A limited-contract system is not for the faint of heart. "There are risks," says Sydney Brenner, a Nobel-prizewinning molecular biologist and senior resident fellow at Janelia, who was once a senior researcher at the LMB. He notes that doses of uncertainty are par for the course. "But if you're passionate enough about doing science, and you have confidence in yourself, you'll be willing to take them," he adds.

The pressures of such models are clear. Working at Bell "was an incredibly highly

competitive atmosphere", says Cherry Murray, a physicist who spent 26 years at the lab in research and management positions, including research vice-president, and is now dean of the Harvard School of Engineering and Applied Sciences in Cambridge, Massachusetts. "You were given some leeway, say for a few years after your arrival, to build up your research programme," she says. But those who consistently stayed in the bottom 10% after that — who weren't exploring imaginative, original ideas as assessed by their managers, and whose research never led to an invention or the possibility of one — were politely asked to leave. Evelyn Hu, an electrical engineer at Harvard who spent nine years as a Bell researcher, recalls a chilling prophecy from company management early on. "I remember attending an orientation for new hires and being told, 'Look to your right, look to your left — in five years, only one of you will be here'," she says.

Those willing to embrace the pressure may face other constraints. The small size of labs in limited-contract institutes can be inhibiting, says Chris Field, director of global ecology at Carnegie and a biologist and environmental Earth systems scientist at Stanford University in California, where he conducts his research but gets no financial or other benefits. "There are some people for whom Carnegie becomes a stage that's not the right size," he says. "Some people find that as they move through their programme, they're more interested in building a bigger lab group."

E. GRINNELL/HARVARD UNIV.

"It was an incredibly highly competitive atmosphere."
Cherry Murray

Those running small labs can risk losing a critical mass of personnel, says Douglas Koshland, a geneticist who spent a long time at Carnegie but accepted a tenured position at the University of California, Berkeley, last year. "If you have four people and two leave, then you've got two left, and that can be painful," he says. But Koshland is still a proponent of small labs, pointing out that the same reduced lab size also enables principal investigators to actually do research, rather than just supervise a dozen or more junior researchers.

Jim Broach is a molecular biologist at Princeton University in New Jersey, but he began his career at Cold Spring Harbor. It was lack of teaching, not of tenure, that drove him into academia. "Postdocs aren't as eager to explore new ideas as graduate students," he says, noting that Cold Spring Harbor does now have an on-campus graduate programme, the Watson School of Biological Sciences, founded in 1999. "Teaching benefits your research — you learn to formulate your questions more precisely and you learn how to organize and present your ideas in a very powerful way," he says.

SOFT LANDING

Being asked to leave a place such as Janelia does not usually spell disaster. Murray notes that any researcher who, voluntarily or otherwise, left Bell while she was there had no problem finding an industrial or tenured academic research position elsewhere. For some, that is a fair exchange. Joanna Aizenberg, a materials scientist at Harvard, spent nine years at Bell, where she loved her work. But when the company began to move away from a basic-research focus to concentrate more on applied, product-driven research, she decided to resign. Shortly after Aizenberg left the company in 2007, she accepted an offer at Harvard. "It's obviously wonderful to have tenure," says Aizenberg, "and to think that whatever happens, I have it."

At Janelia, group leaders who don't receive a renewal offer for a second term will get transitional funding of up to US$1 million a year for two years, a bonus that significantly boosts their recruitment value. Those who get a renewal offer but decide to leave anyway can take their HHMI investigator status, and they get the same transitional funding. "You show up with a really big cheque in your pocket — that's really valuable in academia," says Tim Harris, director of the applied physics and instrumentation group at Janelia. At the LMB, those who are asked to leave are given a month's pay for each year they've worked at the Medical Research Council, up to a maximum of 21 months, and get about a year's notice before they actually have to leave. At Cold Spring Harbor, researchers are reviewed four years into their five-year contracts, so if they are asked to leave, they still have a year to find a job, and may have some money left over from their start-up packages. At Carnegie, departures are often based on mutual agreement. Scientists who go elsewhere receive a lump sum representing their unused annual leave.

Supporters of the short-term model note that tenured academic positions are tough to find — and, in any case, few jobs have long-term guarantees. "Having any job in research, especially now, is such a gift," says Hess. He says researchers should focus on their innovations, rather than on how long their jobs will last. "For me, the reward has always been on the positive side — what's exciting, what's new, and to not be fear-driven about when my job might end," he says. "It's really a blessing to have this kind of opportunity — where people pay you to do what you love doing." ■

Karen Kaplan *is assistant Careers editor at* Nature.

UNIVERSITY SCORING

Movement in the ranks

The California Institute of Technology in Pasadena ranked first for physical sciences in the 2011–12 World University Rankings for subject areas, released last week by *Times Higher Education* (*THE*) in London. Harvard University in Cambridge, Massachusetts, topped the list for life sciences. Changes to the criteria, including a longer collection period for citations, contributed to differences from last year: Pierre and Marie Curie University in Paris rose from 191st place to 30th in physical sciences, and Wageningen University and Research Center in the Netherlands soared from 166th to 17th in life sciences. The *THE* based its findings on indicators in five weighted groups, including research, citations per paper, teaching performance and international engagement. US institutions dominate, but Phil Baty, deputy editor of the *THE*, predicted that China, with its increasing science investments, will soon have a greater presence.

NETWORKING

Mentors wanted

A US science-mentoring service is seeking more advisers after a surge in demand. Since launching an enrolment campaign on Facebook, LinkedIn and Twitter in September, MentorNet, a non-profit group in Santa Clara, California, has signed up 320 graduate and undergraduate students, largely women and mainly from minorities, who want mentors in research and industry. Since 1997, MentorNet has made 30,000 connections, using grants and fees from about 100 US universities. The economy and tight university budgets have hindered expansion, but social media have helped to extend the service beyond member universities, says president and chief executive David Porush.

PHD CANDIDATES

Better student stability

More European nations should recognize doctoral students as employees, said Eurodoc, a Brussels group representing PhD candidates in the European Union, in a statement on 1 November. Norway, the Netherlands and Denmark already classify PhD students as professionals, and give them salaries and benefits. Adding stability and security could draw more people to research, says Sverre Lundemo, Eurodoc's mobility coordinator. Eurodoc is discussing the matter with the European Commission, he says.

THE LONELINESS OF THE LONG-DISTANCE PANDA

Bear necessities.

BY JACEY BEDFORD

JACEY

There were no two ways about it: he clunked when he walked. It was the near hind that was causing the problem. He'd noticed it days ago, soon after crossing the Bering Strait, some 85 relentless kilometres of icy salt water. Pad, pad, pad, clunk. Pad, pad, pad, clunk. Yes, definitely the hip. Too much exercise and way too much salt water.

He'd not been made for this kind of travel, just a little gentle shuffling between bamboo groves, avoiding too many ups and downs. He'd heard them talking about him in the lab before he was released. A bear, they'd called him, *Ailuropoda melanoleuca*, black and white cat-foot, a living fossil. They said his closest ursine relative was the spectacled bear of South America, *Tremarctos ornatus*. He'd checked his GPS, still functioning after all this time, and pondered in the manner of his kind. China to South America was a long walk for a panda, but he could do it if he had to.

It had been many years since he'd walked out of the Qinling Mountains in hilly Sichuan Province and set off to find more of his kind, but although he'd walked north though China to Siberia, he'd found no organic creature larger than a cockroach and no one like himself. He didn't know what had happened to his makers, but it must have happened a long time ago. They'd gone the way of the panda at last.

Back in the early twenty-first century, the International Union for Conservation of Nature had declared his species as *endangered* and *conservation reliant*, but hadn't followed through until it was too late. Maybe they'd thought it enough that the giant panda genome had been sequenced in 2009, but it took more than that to keep a whole species viable. It took bamboo and more bamboo and the resources shrank each year.

That's when they'd made him and others like him. Giant pandas, able to live in the wild on fresh air — a sop to national pride. Something for the tourists to gawp at, perfect in all respects, on the outside at least. On the inside, however, their hearts were tiny fusion reactors; their skeletons, foam-metal; their joints, ceramic; and their muscles, skin and fur woven from various polymers.

He could access his internal clock if he really wanted to and find out how long it had been since the others ceased to function. He was the last of his kind, and pretty soon he'd be unable to walk. Pad, pad, pad, clunk! What would happen then? He imagined himself stranded in this land of glaciers. How long would it take for his fuel cell to run down? Too long. He wondered if it was possible for a creature like him to go mad.

Alaska's broken coastline caused him detour after detour and he stopped checking his clock, fearful of seeing years pass while he relentlessly plodded south. Always south. Pad, pad, pad, clunk through Alaska and British Columbia until the pad, pad, pad, clunk became pad, pad, clank, clunk.

He had to skirt an active volcano where Seattle used to be. Its rumbling and occasional belch of acrid smoke set off his alarm systems and the volcanic grit irritated his retractable claw beds, causing him delays for cleaning and maintenance. Pad, pad, clank, clunk became pad, pad, clank-drag and he was grateful when he reached the benign forests of Oregon.

➲ NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

He was resting his rear hind servo when he heard a whine above him and felt the sudden tug of an antigrav hauler. He was whisked up to a laboratory in the sky. Surprise wasn't one of his inbuilt emotions and he knew all about laboratories, even alien ones, so he settled down feeling only mild curiosity. They shot him with a cryo-anaesthetic that might have worked if he'd been organic, or might have killed him, as the aliens, themselves crystalline blobs, seemed to know so little of Earth-based physiology that they couldn't tell the difference between biological and mechanical. At least they didn't seem interested in taking him apart to see how he worked.

He lay awake for the 90 years it took for their ark ship to return home, grateful that the enforced rest gave his depleted nanites time to repair most of the damage to his hip joints. Occasionally he would open one eye and watch the aliens at work, gleaning an understanding of their communications system, a mixture of audible sounds and electronic impulses. They never seemed to figure out his recorded system responses, though, which were in Mandarin.

In the alien zoo they gave him a huge enclosure that exactly replicated the forest they'd found him in, and put up a communication repeater that he deciphered as *Dominant species of Planet 40698-C*. He shrugged and methodically deleted his GPS database to make room for new information, then set off in a southerly direction. Pad, pad, pad, clink.

He found her on the fourth day, staring round-eyed and hopeful from the lower branches of a tall tree. He'd walked up one side of Earth and back down again to find another of his kind, and here she was, halfway round the galaxy in an alien zoo. From her black coat to her lighter eye markings, she was a prime example of the spectacled bear of South America. He dipped his head in acknowledgement and she dipped hers in return. She climbed down her tree and walked towards him shyly. The faint sound of her servos, pad-clunk, pad, pad, drifted towards him and it made his little fusion reactor glad within his breast. ■

Jacey Bedford *lives on the Yorkshire Pennines with her songwriter husband, and is trying to get all the stories in her head down on paper before her brain explodes.*